The Dissertation Committee for Bradley Jon Wagstaff
certifies that this is the approved version of the following dissertation:

Comparative Genomics and Molecular Population Genetics of *Drosophila* Male Reproductive Genes

Committee:

_____
Ulrich G. Mueller, Supervisor

_____
David J. Begun, Co-Supervisor

_____
Michael C. Singer

_____
James J. Bull

_____
Craig R. Linder

Comparative Genomics and Molecular Population Genetics of *Drosophila* Male Reproductive Genes

by

Bradley Jon Wagstaff, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2004

Comparative Genomics and Molecular Population Genetics of *Drosophila* Male Reproductive Genes

Publication No._____

Bradley Jon Wagstaff, Ph.D.

The University of Texas at Austin, 2004

Supervisors: Ulrich G. Mueller and David J. Begun

DNA sequence data from male reproductive genes in numerous taxa have shown that these genes typically evolve more rapidly than other genes, often as a result of directional selection. In the genus *Drosophila*, the rapidly evolving male accessory gland protein genes (*Acps*) of *melanogaster* subgroup flies have contributed to this observation. *Acps* are small proteins that are transferred to females during mating as a major component of the seminal fluid and are considered agents of chemical communication between the sexes. *Acps* are known to contribute to normal ovulation and sperm storage, as well as increase oviposition rates and reduce female receptivity. Thus, *Acps* are considered likely targets of directional selection because of their potential roles in postcopulatory sexual selection and antagonistic coevolution between the sexes.

Outside of *melanogaster* subgroup *Acps*, little is known about the evolutionary biology of male reproductive genes in *Drosophila*. For example, the male testis contains a richly diverse transcriptome but no studies have explored the evolutionary dynamics of a large set of testis-expressed genes. If clear differences in the evolutionary dynamics of different classes of male reproductive genes exist, empirical documentation of these differences will help identify the specific evolutionary forces at work. Additionally, mating systems differ between *Drosophila* species, potentially affecting the evolutionary dynamics of *Acps* across lineages. Comparative analyses of *Acps* from species with different mating systems are needed to address this issue. Finally, if *Acps* are generally rapidly evolving in *Drosophila* species, comparative analyses of orthology and *Acp* gene loss/gain are needed to determine how *Acps* respond to persistent directional selection across lineages.

The data presented here aim to address these questions. Included are polymorphism and divergence data from 56 genes of *Drosophila arizonae* and *D. mojavensis*, *repleta* group species with mating systems that differ dramatically from *melanogaster* subgroup flies. The sample includes 19 *Acps*, 31 testis-expressed genes, and six more evenly expressed genes. Comparative genomics analyses of *D. melanogaster-D. mojavensis* male reproductive genes and *D. melanogaster-D. pseudoobscura Acps* are also presented to address questions of functional conservation across lineages.

Table of Contents

List of Figures ix

List of Tables xi

Chapter 1: Characterization and Comparative Genomics of *Drosophila mojavensis* Male Reproductive Tract EST Libraries 1

Introduction 1

Materials and Methods 4

cDNA Library Construction and Sequencing 4

D. mojavensis Reproductive Tract Library 4

Preliminary Expression Analysis and *D. mojavensis* Testis cDNA Library Production 4

D. mojavensis Genomic Library 5

Characterization of Reproductive Tract Genes 6

BLAST Methodology 7

Quantitative PCR Evaluation of ESTs 8

Quantitative PCR Statistics 11

Nomenclature 13

Results 13

Library Content 13

Library Quality 15

BLAST Analyses 15

D. melanogaster-D. mojavensis Comparison of Orthology 17

Acps 18

Testis-Expressed and *moj*- Genes From Population Genetics Survey 19

Relative Quantification of Expression 23

D. mojavensis Analysis 23

Comparison of *D. melanogaster* and *D. mojavensis* Expression Patterns 25

Discussion 27

Chapter 2: Molecular Population Genetics of *Drosophila arizonae* and *D. mojavensis* Male Reproductive Genes 32

Introduction 32

Materials and Methods 34

Isolation and Characterization of *D. mojavensis* Genes 34

Population Genetic and Molecular Evolution Analyses 36

Results 38

Evidence of *D. m. baja*-*D. m. mojavensis* Population Substructure 38

Levels of Synonymous and Replacement Polymorphism and Divergence 40

Joint Analysis of Polymorphism and Divergence 44

Discussion 47

Chapter 3: Molecular Population Genetics of *Drosophila arizonae* and *D. mojavensis* Accessory Gland Protein Gene Families 53

Introduction 53

Materials and Methods 57

Gene Discovery 57

Organization of Duplicated *Acps* 58

Ka/Ks Estimation and Hypothesis Tests of Adaptive Protein Evolution 59
Results.....59
Evidence of Gene Duplication.....59
Physical Organization of Duplications 61
Polymorphism and Interspecific Divergence of Duplicate *Acps* 62
Paralogous Ka/Ks Ratios 63
Dating Duplications Relative to *D. arizonae*/*D. mojavensis* Speciation 63
Branch-Specific Divergence of Duplicate *Acps* 65
McDonald-Kreitman Tests of Adaptive Evolution.....66
Discussion.....68
Chapter 4: Comparative Genomics of Accessory Gland Protein Genes In *Drosophila melanogaster* and *D. pseudoobscura*.....72
Introduction.....72
Materials and Methods.....75
Computational Analysis.....75
Empirical Methods.....77
Population Genetics 79
Results.....79
Evidence of Gene Presence.....80
Acp26Aa&*Ab* 80
Acp32CD.....82
Acp53Ea and Duplicates.....83
Evidence of Gene Presence Associated with Genomic Rearrangement.....85

Acp62F 85

Acp70A 89

Evidence of gene absence 91

Acp29AB and *lectin-29Ca* 91

Acp33A 92

Acp36DE 93

Acp63F 94

Acp76A 95

Acp95EF 97

Acp98AB 99

Discussion 102

Figures 110

Tables 122

References 154

Vita 165

List of Figures

Figure 1.1. Comparison of replicate quantitative PCR scores 110

Figure 1.2. Alignment of *D. melanogaster* and *D. mojavensis* microsynteny around the *Tes100*/*115* gene region 111

Figure 1.3. Correlation between absolute levels of expression and degree of tissue-specificity 112

Figure 3.1. Phylogeny of *Acp5* duplicate genes.....113

Figure 3.2. Phylogeny of *Acp16* duplicate genes.....113

Figure 4.1. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp26Aa&Ab* gene region.....114

Figure 4.2. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp32CD* gene region 115

Figure 4.3. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp53Ea* gene region 115

Figure 4.4. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp62F* gene region 116

Figure 4.5. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp70A* gene region 116

Figure 4.6. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp29AB* gene region.....117

Figure 4.7. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp33A* gene region 117

Figure 4.8. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp36DE* gene region 118

Figure 4.9. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp63F* gene region 118

Figure 4.10. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp76A* gene region 119

Figure 4.11. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp95EF* gene region.....119

Figure 4.12. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp98AB* gene region.....120

Figure 4.13. Reverse Northern of *D. pseudoobscura* ortholog candidate (i.e., microsyntenic) regions.....121

List of Tables

Table 1.1. EST Distribution of the *D. mojavensis* Male Reproductive Tract cDNA Library....122

Table 1.2. EST Distribution of the *D. mojavensis* Male Testis cDNA Library....124

Table 1.3. BLAST and Conserved Domain Analysis of *D. mojavensis* ESTs....125

Table 1.4. Quantitative PCR Data for *D. mojavensis* and Related *D. melanogaster* Genes....129

Table 2.1. Evidence of Genetic Differentiation Between *D. m. baja* and *D. m. mojavensis*....133

Table 2.2. Polymorphism and Divergence at Individual *Acp*, *Tes*-, and *moj*- Genes....135

Table 2.3. Polymorphism and Divergence of Gene Classes....138

Table 2.4. Polarized *D. arizonae* vs. *D. mojavensis* divergence....139

Table 2.5. Individual Gene McDonald-Kreitman Tests....141

Table 2.6. McDonald-Kreitman Tests for Gene Classes....144

Table 2.7. Polarized McDonald-Kreitman Tests for Gene Classes....145

Table 3.1. Sample and Distribution of Duplicate Genes....146

Table 3.2. Polymorphism and Interspecific (Orthologous) Divergence of Duplicate *Acps*....147

Table 3.3. Intraspecific (Paralogous) Divergence of Duplicate *Acps*....148

Table 3.4. Branch-Specific Divergence of *Acp21* and *Acp27* Duplicate Families....148

Table 3.5. McDonald-Kreitman Tests of Duplicate Gene Pairs....149

Table 4.1. Gene Intron/Exon Structure, Signal Peptide Prediction, and Amino Acid Sequence Identity Between *D. melanogaster* and *D. pseudoobscura Acps*....150

Table 4.2. Accession Nos. and Initiation Codon Positions for *D. pseudoobscura Acp* Orthologs and Microsyntenic Contigs 152

Table 4.3. Silent and Replacement Polymorphism and Divergence for *Acp26Aa* in *D. melanogaster* and *D. pseudoobscura* 153

Table 4.4. McDonald-Kreitman tests of neutral molecular evolution at *Acp26Aa* in *D. melanogaster* and *D. pseudoobscura* 153

Chapter 1: Characterization and Comparative Genomics of *Drosophila mojavensis* Male Reproductive Tract EST Libraries

Introduction

The functional and evolutionary biology of *Drosophila* seminal fluid proteins from the male accessory gland (*Acps*) has been investigated most thoroughly in *D. melanogaster* (see Wolfner 1997, Chapman 2001, and Wolfner 2002 for reviews). However, the diversity of *Drosophila* mating systems (Markow 1996, 2002) presents opportunities for investigating how variation in mating systems and reproductive biology may affect the function and evolution of *Acps*.

Drosophila mojavensis is a cactophilic fly species within the *mulleri* complex of the *repleta* group. As a member of the subgenus *Drosophila*, *D. mojavensis* is approximately 40-60 million years diverged from the *Sophophora* subgenus of *D. melanogaster* (Powell and DeSalle 1995). The reproductive behavior of this desert *Drosophila* species differs greatly from that of *D. melanogaster*. One major difference involves mating frequency. Remating occurs more rapidly and more often in *D. mojavensis* than in *D. melanogaster*. Within 24 hours of an initial mating, 95% of *D. mojavensis* females tend to remate, while only 2% of *D. melanogaster* females remate in this same time period (Markow 1982, 1996). Frequent remating would favor competition between male ejaculates, whereas infrequent remating would be more likely to favor

genotypes successfully obtaining initial access to females (e.g., Markow 2002). Data from *Drosophila* species suggests that there is a positive correlation between high female remating rates and exaggerated ejaculates in the form of either sperm gigantism or excessive ejaculate donation to female somatic cells (Markow 2002). If male-male and/or male-female postcopulatory interactions contribute to the evolution of these traits, it might be expected that the reproductive tract genes of *D. mojavensis* evolve at an accelerated rate relative to those of *D. melanogaster*.

Another major difference between *D. mojavensis* and *D. melanogaster* is the pronounced insemination reaction that occurs in the female reproductive tract of *D. mojavensis* (Patterson 1947). This reaction occurs immediately after mating and manifests itself as a large mass within the vaginal pouch (Patterson 1946). Homogamic matings of *D. mojavensis* generate a mass that acts as a mating barrier and prevents remating for the several hours that it persists (Patterson 1947, Knowles and Markow 2001). Heterogamic matings with its closely related sister species, *D. arizonae*, trigger an exaggerated insemination reaction that is both harder in composition and lasts significantly longer than the respective homogamic matings of either species (Patterson 1947). Live spermatozoa are not necessary to trigger the insemination reaction. Patterson (1947) was able to show this by duplicating typical heterogamic results with sterile F1 hybrid males that were backcrossed to each of the parental species. He went on to speculate that secreted fluids from the male accessory gland might play a part in the expression of this phenotype. *D. melanogaster* shows no evidence of an insemination reaction (Wheeler 1947; Markow and Ankney 1988).

Knowles and Markow (2001) showed that there is significant variation between populations in the insemination reaction mass duration and size in intrapopulation crosses of *D. mojavensis*. Moreover, the temporal trajectory of the mass differed significantly between intra- and interpopulation crosses, with the size and/or duration of the mass larger and longer in interpopulation crosses (Knowles and Markow 2001). *D. mojavensis* also shows significant among-population variation in the correlated traits of male sperm size and female sperm-storage organ length (Pitnick et al. 2003). These data support the idea that properties of ejaculates or ejaculate-female interactions evolve very quickly in *D. mojavensis*. This apparent rapid evolution of *D. mojavensis* postcopulatory phenotypes may be a consequence of antagonistic coevolution between the sexes (Rice 1996, 1998).

Our interest in the evolutionary genetics of male reproduction related proteins in *Drosophila* motivated the isolation, molecular characterization, and evolutionary and population genetics investigation of *Acps* and testis-expressed genes in *D. mojavensis* and its sister species, *D. arizonae*. Here we present our isolation and characterization of genes derived from *D. mojavensis* cDNA libraries.

Materials and Methods

cDNA Library Construction and Sequencing

D. mojavensis Reproductive Tract Library

PolyA-enriched mRNA was prepared with the MicroPolyA-Pure kit (Ambion, Austin, TX) from 50 whole reproductive tracts of male *D. mojavensis* flies. First strand cDNA was reverse-transcribed with the SMART PCR cDNA synthesis system reagents and protocol (CLONTECH Laboratories). Second strand product was produced with the Expand High-fidelity polymerase system (Roche Molecular Biochemicals). Cycling parameters were programmed as instructed by the manufacturer, including a 4-minute extension step for 10 total cycles. The second strand product was cloned into the TOPO vector (Invitrogen, USA) and used for bacterial transformations according to the manufacturer's instructions. Colony PCR was carried out using cloning-vector-derived primers (M13 reverse and T7) on 480 colonies (i.e., five 96-well plates). The resulting PCR products were purified prior to sequencing with M13R and T7 primers on an Applied Biosystems 377 automated sequencer (ABI, USA). These sequences included 53 unique transcripts.

Preliminary Expression Analysis and *D. mojavensis* Testis cDNA Library Production

Dot blots prepared from PCR products of these 53 distinct clones were hybridized separately to ^{32}P-labeled cDNAs derived from *D. mojavensis* accessory glands and testes.

Hybridizations were carried out at 65^{o}C in a buffer consisting of 0.5M NaPi (pH 7.2), 7% SDS, 1mM EDTA. Filters were washed at 60^{o}C with buffer at 40mM NaPi, 1% SDS, and 1mM EDTA. These hybridizations suggested that the majority of the clones represented accessory gland transcripts rather than testis transcripts.

To increase the sample size of testis-expressed genes we made a testes cDNA library. This library was produced as described above for whole reproductive tracts, but with 50 *D. mojavensis* dissected testes as the source tissue. This library was sequenced to the point of producing 119 unique ESTs.

D. mojavensis Genomic Library

D. mojavensis genomic DNA was partially digested with *Sau3A* and size-fractionated by electrophoresis through a 0.6% agarose gel. DNA fragments between 9 and 23 kb were selected via gel extraction (Qiagen), ligated to Lambda DASH II/*BamHI* vector (Stratagene), and packaged using the Lambda DASH II/Gigapack II Cloning kit (Stratagene). The resultant library consisted of approximately 2.3 x 10^{6} plaque forming units. Plaques were screened with ^{32}P-labeled *D. mojavensis* target DNA. Lambda DNA was purified from selected plaques and *D. mojavensis* genomic inserts were amplified using T3/T7 vector primers and LA-Taq long PCR polymerase (TaKaRa, Japan). The resulting PCR products were sheared by sonication and the fragments were blunt-ended using Klenow fragment of DNA polymerase and T4 DNA polymerase. Fragments of 1-2

kb were isolated from a low-melting agarose electrophoresis gel and cloned into the pUC18/SmaI/BAP vector with a Ready-to-Go kit (Amerisham Biosciences, Piscataway, NJ). Sequencing was performed on an ABI Prism 3700 sequencer through 7X coverage. Consensus sequences were assembled using the SeqMan program of the DNASTAR software package (Lasergene, Madison, WI).

Characterization of Reproductive Tract Genes

A subset of genes isolated from both libraries was selected for possible population genetic analysis (see Chapter 2) and was scrutinized in more detail. Each clone sequence was subjected to an open reading frame (ORF) analysis by the GeneJockey software program (Biosoft, Inc., Ferguson, MO). If a putative initiation codon followed by an ORF covering at least 70% of the EST could not be identified, we used RACE to gather additional cDNA sequence data. Reproductively mature *D. mojavensis* adults of both sexes served as the tissue source for the RACE-ready template. mRNA was isolated using the MicroPolyA-Pure kit (Ambion, Austin, TX). RACE-ready cDNA was prepared and target molecules were PCR amplified and isolated using the GeneRacer (Invitrogen) kit according to the manufacturers instructions. The protocol separates the truncated from the complete and mature mRNA products, preferentially selecting the full-length transcripts for first-strand cDNA synthesis. Thus, RACE products derived from such a library should provide high quality information on the 5' ends of transcripts. Overall, several criteria were used to identify the set of ORFs ultimately used in molecular

evolutionary analysis: (i) size and position of candidate ORFs within an EST, (ii) presence of an apparent signal peptide sequence for putative *Acps* (Bendtsen et al. 2004), (iii) tBLASTx homology to orthologous genes in public databases, (iv) presence/absence of indel mutations and possible frameshift consequences in polymorphism data from genomic DNA (from Chapter 2). Only strongly supported ORFs were used in evolutionary analysis.

BLAST Methodology

All unique ESTs were compared to *D. melanogaster* through a pre-determined pipeline of BLAST analyses to one or more FlyBase Release 3.1 databases (Altschul et al. 1997). Default BLAST parameters were used except that the cutoff value for significance was set to E = 0.01. The pipeline started with BLASTp (protein to predicted *D. melanogaster* proteins) queries of all ESTs for which an ORF was well established. ESTs that returned highly significant (E < 1e-8) *D. melanogaster* sequences were not queried further. The remaining ESTs were BLASTx (nucleotide to protein) queried to the same *D. melanogaster* database. Once again, ESTs that returned highly significant sequences were not queried further. This pipeline continued through tBLASTx (nucleotide to nucleotide query, using all six possible protein translations of the sequences) and BLASTn (nucleotide to nucleotide) queries of predicted *D. melanogaster* genes and chromosome arms. For the ESTs that returned no *D. melanogaster* sequences at E < 0.0001, the NCBI wgs (whole genome shotgun) database was tBLASTx queried

with the same default parameters (Altschul et al. 1997). The NCBI wgs database includes many complete insect genomes, including *D. pseudoobscura* and the mosquito, *Anopheles gambiae*. All *D. mojavensis* ESTs were also tBLASTx or BLASTn queried (BLASTn was only used if tBLASTx failed to return sequences of $E < 0.0001$) to the *D. melanogaster* dbEST database using default BLAST parameters and an E score cutoff of 0.01. Finally, we queried the NCBI CDD server (Marchler-Bauer et al. 2003) for all EST sequences with known protein sequences.

Quantitative PCR Evaluation of ESTs

Genes subjected to population genetic analyses were characterized by quantitative PCR for expression heterogeneity across tissues. For the subset of genes in which a related *D. melanogaster* gene was identified, quantitative PCR was also carried out in *D. melanogaster* to provide comparisons of expression between lineages. The purpose of this analysis was to assign genes to three expression classes: *Acp*, testis-expressed, and other.

A total of 80 *D. mojavensis* and 40 *D. melanogaster* male flies were used in tissue dissections. All flies were reproductively mature and were dissected in RNA*later* (Ambion) into three tissue categories: accessory glands, testes, and carcasses without the reproductive tracts. The tissues from each of these preps were then divided equally into two replicate samples for RNA isolation. Likewise, 40 whole reproductively mature female flies from each of these species were evenly split into replicate 20 fly RNA preps.

Total RNA from all tissues was extracted using Trizol Reagent (Invitrogen) according to the manufacturer's specifications. Total RNAs were then purified through Rneasy (Qiagen) columns and treated with DNase according to manufacturers instructions (Qiagen). The purified RNAs were then reverse transcribed at a concentration of 20ng/ul using TaqMan RT (reverse transcription) reagents (Applied Biosystems). These first-strand cDNAs served as the templates for quantitative PCR analysis.

Quantitative PCR was performed using an ABI Prism 7700 Sequence Detector and SYBR green PCR core reagents (Applied Biosystems). Amplification primers were designed with Primer Express (Applied Biosystems). For every 20ul PCR reaction, 0.5ul of first strand cDNA was used. Quantitative PCR conditions were 94^0C for 10 min followed by 40 cycles of 94^0C for 20 s, 59^0C for 30 s, 72^0C for 30 s. In order to insure that only a single amplicon was produced in each reaction, a dissociation step was added to the end of the run according to manufacturer's instructions. All primer pairs produced a single product. A total of 13 quantitative PCR reactions were processed for each gene. Three reactions were run for each of the four tissues: one for each of the two replicate RT reactions as well as a single -RT reaction derived by drawing equally from the -RT templates of the replicates. The 13^{th} reaction was a no-template control. There were no signs of genomic contamination or primer x reagent interactions in any of the reactions.

Quantitation of the data followed the $2^{-\Delta\Delta C_T}$ methods of Livak and Schmittgen (2001). Quantitative PCR analysis uses the ability to quantify double-stranded product during amplification to estimate C_T, the cycle at which amplified product exceeds a pre-determined threshold. To control for different first-strand cDNA concentrations across

templates, as well as run and reagent effects, our ΔC_T scores were calculated by subtracting experimental gene C_T scores from housekeeping gene C_T scores derived from the same tissue and experimental plate. Therefore, the more abundant a particular transcript is, the lower its measured ΔC_T score. Since ΔC_T scores serve as useful approximations of transcription levels, they can be loosely compared across genes. The housekeeping control for both species was the ribosomal protein, *CG7808*. This gene was identified in the original *D. mojavensis* reproductive tract cDNA library (*moj12*) and is highly conserved between species (96% protein similarity).

Our calculation of $2^{-\Delta\Delta C_T}$ reflects fold change in gene expression of the most abundant tissue template (lowest ΔC_T score) relative to the second most abundant tissue template for any given gene. This approach, as opposed to scoring relative to the third or fourth most abundant tissue template, minimizes fold difference values, thus providing conservative lower-bound estimates for actual differences between transcriptome profiles of these tissues. Furthermore, there were several instances in which quantitative PCR product was only detected in two of the four templates. The two replicate $2^{-\Delta\Delta C_T}$ scores for each gene were always independently calculated and then averaged for the reported values.

Quantitative PCR Statistics

Since we derived two independent $2^{-\Delta\Delta C_T}$ scores for every gene and each of the four templates, the similarity of replicate pair scores can be used to determine the amount of experimental error. First, our graph of replicate ΔC_T scores for the most abundant tissue of each surveyed gene (n = 91, Figure 1.1) shows a high degree of similarity between replicate pairs. The best-fit line closely matches the data (R^2 = 0.979). Furthermore, the slope of this line (m = 0.985) is very close to the slope that would be expected (m = 1) if all replicate ΔC_T pairs were identical. These results indicate that our measurements of ΔC_T are highly repeatable.

We used our replicate $2^{-\Delta\Delta C_T}$ scores to determine threshold fold differences that are sufficiently disparate to represent significant differences between scores. To approximate a gamma distribution, we calculated ratios of replicate pairs by dividing the higher $2^{-\Delta\Delta C_T}$ score by its counterpart, and then subtracted one. A total of 91 replicate reaction pairs generated a distribution ranging from zero to 18.24. We then use the x_0 value at which the area under the frequency distribution ($0 \leq x \leq x_0$) is equal to 0.95 to establish the critical threshold for significant differences between $2^{-\Delta\Delta C_T}$ scores. For the complete data set, $2^{-\Delta\Delta C_T}$ scores > 7.84 represent significant differences between tissues ($P < 0.05$). This is a very conservative critical threshold estimate because genes that are highly tissue-specific (those with high $2^{-\Delta\Delta C_T}$ scores), and leave no doubt of expression

differences between tissues, are bound to widen replication error. Since fold difference is calibrated relative to the second most abundant tissue (second lowest ΔC_T score), cases of highly tissue-specific expression involve a calibrator that must be present only in trace amounts. Sampling error that is likely to occur from quantifying C_T scores from these trace templates will carry over to $2^{-\Delta\Delta C_T}$ scores of highly tissue-specific genes. Many of our genes have very high $2^{-\Delta\Delta C_T}$ scores (see Table 1.4), an indication of high tissue-specificity. Restricting our statistical analysis to genes with $2^{-\Delta\Delta C_T} < 50$ ($n = 28$), the critical threshold for significance is reduced to 3.25 ($P < 0.05$). Further narrowing the analysis to genes with $2^{-\Delta\Delta C_T} < 15$ ($n = 24$) reduces the critical threshold to 2.10 ($P < 0.05$).

The different critical threshold values for different subsets of the data demonstrate that the highest $2^{-\Delta\Delta C_T}$ scores account for much of the replication error. Therefore, we view the critical threshold of 2.10 as most informative because it is derived from the very data whose expression patterns are most in doubt. Even so, we choose a conservative critical threshold of $2^{-\Delta\Delta C_T} = 5.0$ for the purpose of categorizing genes as either *Acps* or testis-expressed. Though it is somewhat arbitrary to choose this specific value for our tissue-specificity threshold, categorization of genes would not change dramatically by choosing a more conservative threshold. For example, a critical threshold of 18 would only re-categorize three *Tes*- genes as *moj*- genes.

Nomenclature

Unique ESTs were given numbered names as they were sequenced (1-53 for reproductive tract library ESTs, 100-218 for testis library ESTs). Prefixes for numbered EST names were added according to expression patterns, with *Acp*- preceding accessory gland genes, and *Tes*- preceding testis-expressed genes. Genes from the quantitative PCR analysis showing at least five-fold greater expression ($2^{-\Delta\Delta C_T} > 5$) in either accessory gland or testis were categorized as *Acps* and testis-expressed genes, respectively. Those genes that did not exceed this threshold (*moj9*, *moj29*, *moj30*, *moj32*, *moj137*, and *moj152*) were given the *moj*- prefix to avoid a connotation of tissue-specificity. Four *Acps* (*Acp5*, *Acp16*, *Acp21*, and *Acp27*) are members of recently duplicated gene families and are given an additional -*a* or -*b* suffix to differentiate between members (not all recent duplicates are mentioned here; these families are covered in greater detail in chapter 3). An additional five genes (*Acp4*, *Acp15*, *Acp17*, *Acp23* and *Acp36*) were clearly *Acps* based on our dot blot data. The remaining ESTs were simply given the *moj*- prefix, as not enough was known about their expression patterns to be more specific.

Results

Library Content and Quality

Table 1.1 shows the numbers of sequenced clones corresponding to each unique transcript from the *D. mojavensis* male reproductive tract cDNA library. Minimal

sequencing revealed that most of the transcripts corresponded to just a few genes. Of the first 139 successfully sequenced clones, 35 corresponded to *Acp1*, 27 to *Acp5*, and 18 to *Acp17*. This group of 139 clones also included 13 singletons and 10 transcripts represented by 2-9 clones each. In order to identify unique ESTs more efficiently, additional clones were screened by multiplexed PCR reactions that included primer pairs specific to *Acp1*, *Acp5* and *Acp17*. Clones not corresponding to any of these three genes were then sequenced. Although this multiplex PCR strategy was not 100% efficient, we were able to identify an additional 27 unique ESTs from only 60 additional sequencing reactions. In total, 53 unique ESTs were revealed. The average length of all 199 ESTs was 438 bp.

The *D. mojavensis* male testis cDNA library was constructed to identify more testis-enriched transcripts because dot blot data revealed a strong accessory gland bias in the first library. Table 1.2 summarizes the EST content for this library. The distribution of replicate ESTs differs dramatically from the original reproductive tract library. The testis library has a much higher complexity than the reproductive tract library, with 105 of 162 clones present as single copy sequences. Similarly high complexity of a testis cDNA library was previously observed in *D. melanogaster* (Andrews et al. 2000), suggesting that this might be a general property of the *Drosophila* testis transcriptome. In total, 162 sequencing reactions returned an average EST length of 451 bp and produced 119 unique ESTs. The large proportion of singleton testis-derived ESTs are a clear indication that our view of the *D. mojavensis* testis transcriptome remains far from complete.

Library Quality

The quality of these libraries, as measured by the completeness of the 5' ends of the ESTs, was assessed by two methods on a total of 155 ESTs. First, for transcripts represented by multiple clones, we compared the similarity of 5' ends among clones, with the assumption that multiple, similar 5' ends are more likely to represent the actual 5' end of a gene. If there were at least 5 copies of a particular EST, the longest was assumed to be full-length and was not counted since it served as the quality indicator for other replicates. Second, several transcripts were subjected to 5' RACE verification to determine whether the original clone(s) isolated from the library were complete at the 5' end. In total, 76 of the ESTs were compared against 5' RACE products and the remaining 79 against putative full-length clones. Almost 80% (123) were full-length (within 10 bp of the longest duplicate EST), 15 were within 11-30 bp of full-length, and the remaining 17 were more than 30 bp shorter than the reference longest transcript or RACE product at the 5' end.

BLAST Analyses

Table 1.3 shows the results of the BLAST analyses to *D. melanogaster* (only ESTs with match scores of $E < 0.01$ are listed). None of the ESTs that failed to match *D. melanogaster* sequences matched any other NCBI database sequences. Accessory gland

genes show a much lower level of conservation between species than their testis counterparts. Only 33% (8 of 24) of *Acps* generated significant hits (E < 0.01), compared to 82% (27 of 33) for testis-expressed genes. A 2x2 contingency table is significantly heterogeneous ($P << 0.01$). Furthermore, the median E value of the eight *Acps* with E < 0.01 is 1e-3, a value typically above the threshold used to establish orthology. The median testis-expressed gene E score of 2e-21 is much higher than that observed for *Acps*. According to our quantitative PCR analyses, only six genes (*moj9*, *moj29*, *moj30*, *moj32*, *moj137*, and *moj152*) are more evenly expressed across tissues. Of these six, all had highly significant BLAST matches to *D. melanogaster* sequences, with a median E score of 5e-42. The remaining *moj-* ESTs are closer to the testis-expressed ESTs than *Acps*, with 55% (59 of 108) returning E < 0.01 vs. *D. melanogaster* and a median E score of 1e-27. This is not surprising, given that most *moj-* genes are from the testis EST library and have not been tested for tissue-specificity.

Of the 27 *D. mojavensis* testis ESTs that appear to have related *D. melanogaster* genes, 20 have BLAST hits to the *D. melanogaster* testis EST collection (Andrews et al. 2000), suggesting testis expression patterns between species are generally conserved. With the exception of the top *D. melanogaster* BLAST match to *Tes118*, for which we have no *D. melanogaster* expression data, our expression analysis from a later section (see Table 1.4) shows that the remaining six primary *D. melanogaster* BLAST matches are also testis-expressed genes despite their absence from the *D. melanogaster* testis EST collection.

D. melanogaster-D. mojavensis Comparison of Orthology

Large gene families and shared protein domains lower BLAST E scores and obscure inferences of gene genealogies between *D. melanogaster* and *D. mojavensis*. Therefore, very low E scores (e.g., E < 1e-10) are not necessarily indicative of true orthologous gene pairs. Conservation of intron-exon structure is expected for genes of shared ancestry (Meyer and Durbin 2004) but not for unrelated genes that only share protein domains. For example, human-mouse orthologs have the same number of coding exons approximately 86% of the time (Mouse Genome Sequencing Consortium 2002). Thus, gene pairs with conserved intron-exon structure and large E score differences (e.g., E > 1e-10) between primary and secondary BLAST hits are likely to represent true orthologs. Our population genetic data (Chapter 2) allowed us to inspect intron-exon structure for a subset of *D. mojavensis* genes (i.e., genes from Table 1.4). Using this information, along with comparisons of primary and secondary BLAST similarity and protein size, we label putative *D. melanogaster* orthologs for all *Acps* and other genes from the population genetic analysis (indicated by an asterisk, Table 1.3). For the remaining ESTs, strong individual conclusions are not warranted since we do not have genomic sequence data and the lack of an asterisk should not be taken as evidence against orthology. However, large differences between primary and secondary E scores and the lack of conserved domains for many ESTs leave little doubt that several of these primary BLAST hits represent true *D. melanogaster* orthologs. Below, we discuss our conclusions on *D. melanogaster-D. mojavensis* relationships for all *Acps* and other genes

from the population genetics survey. We note, however, that orthology cannot be definitively proven with these data.

Acps

Of the eight *Acps* that show BLAST similarity to *D. melanogaster* genes, only *Acp36* and *CG16713* (Table 1.3) represent a potential orthologous pair. Though we do not have genomic sequence data for *Acp36*, other evidence supports orthology. *Acp36* and *CG16713* both consist of 82 residues and share a Kunitz domain that covers 59 of those residues. A protein alignment is 57.3% identical (47/82) and contains no gaps. The alignment outside of the domain is still 34.8% identical (8/23). Though another protein with a Kunitz domain (*CG16712*) generates a highly significant secondary BLAST hit, a protein distance tree clusters *Acp36* with *CG16713*. We cannot definitively state that these genes are true orthologs. However, *Acp36* and *CG16713* represent the most likely example from these data of *D. melanogaster-D. mojavensis Acp* orthology.

Three more *Acps* (*Acp1*, *Acp2*, and *Acp25*) are part of a gene family and appear to be paralogous to the *Acp53-* gene family in *D. melanogaster* (Table 1.3). There are four known members of the *Acp53-* gene family, all related through tandem duplication (Holloway and Begun 2004). A protein distance tree clusters the three *D. mojavensis* genes, rather than generating three interspecific pairs as would be expected under a hypothesis of orthology. The *Acp1* protein is 45.8% similar to *Acp25* and 34.7% similar to *Acp2*. *Acp2* and *Acp25* are only 30.5% similar. An alignment of *Acp2* and

Acp53C14a generates the best interspecific pairing, at only 23.7% similarity. Thus, it is likely that the *D. melanogaster Acp53-* duplicate genes represent paralogous counterparts to the *Acp1-2-25 D. mojavensis* gene family.

The remaining *Acps* do not present compelling cases for orthology for several reasons. *Acp19*, along with *Tes33*/*moj49*/*Tes104*, is part of a large family of SCP-related genes (Table 1.3). A protein distance tree with these four *D. mojavensis* genes and the six closest (lowest E scores) *D. melanogaster* SCP-related genes fails to produce any interspecific gene pairs, thus providing no hint of orthology. Next, *Acp4* and *CG11395* share no protein domains but have a fairly low BLAST E score (E = 8e-08). However, these proteins are dramatically different in size with *Acp4* at 119 residues and *CG11395* at 456 residues. *Acp48* and *Spn43Aa* are probably not orthologous either. Differences in intron-exon structure and very low BLAST similarity (E = 2e-03) combined with shared Serpin domains suggest orthology is unlikely. Finally, the very low BLAST similarity (E = 5e-03) between *Acp27* and *Def* does not provide enough evidence to conclude that these genes represent an orthologous pair.

Testis-Expressed and *moj-* Genes From Population Genetics Survey

Most testis-expressed and *moj-* genes from the population genetics survey have clear *D. melanogaster* orthologs (Table 1.3). As stated above, *Tes33* and *Tes104* are part of an SCP-related gene family and have no obvious orthologous counterparts. Three additional testis-expressed genes, *Tes114*, *Tes120*, and *Tes123*, are also part of gene

families that obscure interspecific relationships. Two remaining genes, *Tes101* and *Tes109*, are too dissimilar to their *D. melanogaster* counterparts (E = 6e-03 and E = 7e-04, respectively) to conclude that they represent orthologous pairs. All *moj-* genes and the remaining testis-expressed genes from the population genetics survey generate primary BLAST hits to their putative *D. melanogaster* orthologs. For most of these genes, large disparities between primary and secondary BLAST matches provide the best evidence of orthology. Population genetics data provided further evidence, demonstrating conserved intron-exon structure for all putative orthologous pairs.

Two genes, *Tes14* and *Tes118*, appear to be orthologous to unannotated *D. melanogaster* genes and warrant additional comment. tBLASTn analysis of *Tes14* to predicted *D. melanogaster* genes produces a significant hit (E = 1e-25) to *CG8446*. However, the match is not to *CG8446* CDS. Instead, the putative termination codon of the unannotated match is separated from the initiation codon of *CG8446* by 978 bp. Our protein alignment suggested a *D. melanogaster* ortholog with two coding exons of identical size to *D. mojavensis Tes14*, separated by a single 89 bp intron. Furthermore, the corresponding proteins in both species are 80 residues in length and the alignment is 68.8% identical with no gaps. However, this putative CDS in *D. melanogaster* includes intron splice sites that are not consistent with *CG8446* mRNA, suggesting a different mRNA species must be responsible for this unannotated *D. melanogaster* gene. To be precise, our predicted *D. melanogaster* ortholog CDS is a subset of the *CG8446* mRNA, except for an included 26 bp sequence that is part of an intron that is spliced out of *CG8446* mRNA. In fact, 5' RACE of *D. melanogaster* cDNA proves this exact splice

exists, as suggested by our computational analysis. Therefore, we believe the *D. melanogaster* sequence in question is an unannotated gene and is orthologous to *Tes14*.

Tes118 shows evidence of orthology to unannotated *D. melanogaster* sequence that is not part of the current genome assembly (Release 3.1). The "all *Drosophila* sequences" database retuned a significant tBLASTn match (E = 3e-12) to a chromosome 2 clone (ACO16129) while BLAST searches of predicted genes and chromosome arms returned no significant results. The tBLASTn match covered two distinct sections of protein sequence. The first tBLASTn hit covered residues 7-96 of our *D. mojavensis Tes118* protein sequence (E = 3e-12, 49.4% protein similarity). The second tBLASTn hit covered residues 120-312 (E = 5e-11, 37.9% protein similarity). There are 134 bp in *D. melanogaster* between tBLASTn hits, showing preserved microsynteny with the 7-96 residue match 5' of the 120-312 residue sequence. However, there are in-frame stop codons within this sequence, indicating the likely presence of an intron. We do not have *D. mojavensis* population genetics data covering the genomic sequence with the putative intron. However, the dual, microsyntenous tBLASTn matches are a strong indication of *D. melanogaster*-*D. mojavensis* orthology.

The lack of BLAST similarity of many *D. mojavensis* genes to *D. melanogaster* sequences does not preclude orthology. We were able to use *D. mojavensis* genomic sequence data from our phage library to identify a *Tes100*/*Tes115* ortholog in *D. melanogaster*. *Tes100* and *Tes115* were the two most frequently sequenced ESTs in our testis cDNA library (Table 1.2) and maintain no BLAST similarity to *D. melanogaster* sequence databases. Our genomic sequence data covers 19.6 kb and is derived from a

phage clone that was detected via hybridization to ^{32}P-labeled *Tes100* genomic DNA. We confirmed the presence of *Tes100* within this sequence. Moreover, our *Tes115* gene was also detected with 1422 bp of intergenic sequence separating the termination codon of *Tes100* from the initiation codon of *Tes115*. *Tes100* and *Tes115* contain two coding exons, with the CDS of the first exon covering 58 bp in each gene. The protein sequences are 50% identical (28/56) across alignable residues. *Tes100* and *Tes115* proteins are 57 and 69 residues in length, respectively. Given the conservation of intron-exon structure, their tandem positioning, and protein similarity, we conclude that *Tes100* and *Tes115* are related through tandem duplication.

The 19.6 kb sequence also contains four *Tes100*/*Tes115* flanking genes that were identified by BLAST analysis to *D. melanogaster*. Covering the immediate 8 kb 5' of *Tes100*, BLASTp matches were (from 5' to 3') to *CG8019* (E = 0.0), *CG6502* (E = 0.0), and *CG8009* (E = 7e-58). The remaining BLASTp match was to *CG6491* (E = 2e-91), 1.3 kb 3' of *Tes115*. *D. melanogaster-D. mojavensis* microsynteny is preserved for these genes, with part of the microsyntenic region illustrated in Figure 1.2. The candidate region for *D. melanogaster* orthology, between *CG8009* and *CG6491*, is less than 2 kb and contains a single annotated gene, *CG18628*. Like *Tes100* and *Tes115*, *CG18628* CDS contains two exons, with the first at 58 bp in length. *CG18628* protein contains 63 residues and is of intermediate length relative to *Tes100* and *Tes115* proteins. Protein alignments of *Tes100* and *CG18628* are 33.9% similar while *Tes115* and *CG18628* are only 25.4% similar. Thus, our data suggest that the *Tes100*/*Tes115* duplication occurred

subsequent to the *D. melanogaster-D. mojavensis* lineage split and that *CG18628* is orthologous to the common ancestor of *Tes100*/*Tes115*.

Relative Quantification of Expression

D. mojavensis Analysis

Table 1.4 summarizes the quantification results for all *D. mojavensis* genes surveyed, as well as several *D. melanogaster* genes that will be discussed in the next section. As stated above, these results served to name genes according to tissue-specificity. Of the 58 total *D. mojavensis* genes selected for quantitative PCR, 19 are primarily expressed in the accessory glands, 33 are primarily expressed in the testis, and the remaining six (*moj9*, *moj29*, *moj30*, *moj32*, *moj137* and *moj152*) are more evenly expressed, as indicated by $2^{-\Delta\Delta C_T} < 5$. The vast majority of the 58 genes appear to be either tissue-specific or highly tissue-enriched in expression, with 46 out of 58 genes being at least 50 times more abundant in one tissue than any other. These data are consistent with the observation that *Acps* are typically secreted peptides (Wolfner 1997). All 19 *Acps* contain putative signal peptide sequences (designated by an * in Table 1.4). In contrast, only three of six *moj-* genes and five of 33 *Tes-* genes contain putative signal peptide sequences.

Figure 1.3 depicts the relationship between ΔC_T and $2^{-\Delta\Delta C_T}$ scores. The more tissue specific genes (high $2^{-\Delta\Delta C_T}$ scores) tend to have higher absolute levels of expression (lower ΔC_T) ($r = -0.5$, $p = 0.0002$). Also, our ΔC_T scores suggest that the six

most abundant genes are all *Acps*. This is not surprising, given our observation of an *Acp* bias from clone sequences of our complete male reproductive tract cDNA library and the low EST diversity of the reproductive tract library relative to the testis cDNA library (Tables 1.1-2). In contrast, our $2^{-\Delta\Delta C_T}$ scores suggest that the 19 most tissue-specific genes are all expressed in the testis. This could be a true difference between accessory gland and testis transcriptomes. However, we believe that this observation is an artifact of trace accessory gland contamination in testis tissue preps, leading to an artificial plateauing of accessory gland gene $2^{-\Delta\Delta C_T}$ scores. Due to the transparent and fragile nature of accessory gland tissue, this type of contamination is much more likely than the converse. In fact, all *Acps* are scored relative to ΔC_T scores of testis templates rather than females or remaining male carcasses (see tissue order, Table 1.4). However, low levels of this one-way contamination should not dramatically affect our conclusions of tissue-specificity. Since contamination would decrease $2^{-\Delta\Delta C_T}$ scores for every *Acp* gene proportionally, the fact that several *Acps* clearly show very large fold differences means this trace contamination is very small. For example, *Acp2* ranks as the most tissue-specific *Acp*, scored at 933 times more abundant in accessory glands than the testis (Table 1.4). Conservatively assuming it is not transcribed in the testis, this roughly means that there are 933 parts accessory gland material in the accessory gland tissue prep for every one part of contaminating accessory gland material in the testis tissue prep. Thus, we do not conclude, for example, that *Tes101* ($2^{-\Delta\Delta C_T} = 36656$) is more tissue-specific than *Acp2* ($2^{-\Delta\Delta C_T} = 933$). On the other hand, *Acp2* is certainly more tissue-

specific than *Acp25* ($2^{-\Delta\Delta C_T} = 51$) since contamination would affect each *Acp* gene $2^{-\Delta\Delta C_T}$ score in a similar manner.

Comparison of *D. melanogaster* and *D. mojavensis* Expression Patterns

Expression patterns of *D. melanogaster* genes reported in Table 1.4 include putative orthologs as well as other possibly related genes and genes with shared domains. First, two *D. melanogaster* genes, *CG1385* and *CG14926*, are included in this analysis despite their low BLAST similarity (E > 1e-03) to *Acp27a* and *Tes101*, respectively. Unlike *Acp27a*, *CG1385* (*Defensin*) is primarily expressed in male non-reproductive tissues (Table 1.4). *CG1385* is a *Drosophila* immune system protein involved in antibacterial defense (Dimarcq et al. 1994). If $2^{-\Delta\Delta C_T}$ was calculated relative to accessory gland rather than female tissue, *CG1385* fold difference would be 145 rather than 2.82. Clearly *CG1385* is not an accessory gland gene in *D. melanogaster* and we have no reason to suspect orthology, though the protein similarity leaves unanswered questions about the possibility of functional convergence. *Tes101* and *CG14926* present a different situation. Despite their limited BLAST similarity, they are both highly testis-specific in expression. Our population genetic data for *Tes101* covers one intron and is consistent with the intron-exon structure of the alignable portion of *CG14926*. Thus, it is still possible that these genes are orthologous, albeit with significant sequence divergence.

We do not expect gene expression patterns to be similar for all members of a gene family. Three of our *D. mojavensis*-*D. melanogaster* comparisons (*Acp19*-*CG9538*,

Acp48-*CG12172*, and *Tes33*-*CG5106*) involve genes from large families with shared domains and no evidence of orthology. Only one pair, *Tes33*-*CG5106*, shows similar, testis-specific expression patterns. Our remaining gene family comparison is smaller and contains no conserved domains. The *D. melanogaster Acp53*- gene family only has four known duplicated members (Holloway and Begun 2004) and no shared protein domains. As stated previously, a protein distance tree generates intraspecific gene clusters, suggesting one-to-one orthologous pairs are unlikely. Thus, it is no more correct to pair our *D. melanogaster Acp53Ea* expression analysis with *Acp25* (Table 1.4) than with either *Acp1* or *Acp2*. Just as all three *D. mojavensis* genes are clearly *Acps*, *Acp53Ea* shows clear accessory gland expression. The degree of *Acp53Ea* tissue-specificity ($2^{-\Delta\Delta C_T} = 47.5$) is much more similar to *Acp25* ($2^{-\Delta\Delta C_T} = 50.8$) than either *Acp1* ($2^{-\Delta\Delta C_T} = 566$) or *Acp2* ($2^{-\Delta\Delta C_T} = 933$). In fact, there is a significant difference between the $2^{-\Delta\Delta C_T}$ scores of *Acp25*/*Acp53Ea* and *Acp1*/*Acp2* ($P < 0.05$).

The remaining *D. mojavensis*-*D. melanogaster* comparisons involve putative orthologous pairs. *D. melanogaster* counterparts to the six *moj*- genes are likewise roughly evenly expressed across tissues, with *CG3654* generating the highest $2^{-\Delta\Delta C_T}$ score (2.53). At most, the differences between pairs in highest scoring tissue (see tissue order, Table 1.4), might reflect subtle differences between *D. mojavensis*-*D. melanogaster* expression profiles.

All putative orthologs to *D. mojavensis* testis-expressed genes are also testis-expressed in *D. melanogaster*. However, there are variations in degree. At the most extreme, *D. melanogaster CG3708* is approximately 164-fold more testis-specific than

Tes129. There are also large fold differences between *Tes106*/*CG30334* (97-fold), *Tes110*/*CG15219* (24-fold), and *Tes127*/*CG10090* (53-fold). These comparisons reflect significant differences between *D. mojavensis*-*D. melanogaster* expression profiles at these genes ($p < 0.05$). Several additional testis-expressed genes are borderline significant with fold differences greater than five. We also note that the paralogous *Tes100* and *Tes115* genes and their *D. melanogaster* counterpart, *CG18628*, all have similar expression profiles.

A final *D. melanogaster* expression profile is included for *CG8446* (Table 1.4). As explained previously, the peptide sequence of the unannotated *D. melanogaster* ortholog to *Tes14* is derived from an mRNA species that is related to *CG8446* mRNA. These results show that there is a stronger testis-expression bias of the unannotated *D. melanogaster* ortholog than the *CG8446* gene. Even so, *D. mojavensis Tes14* shows a greater degree of testis-specific expression than the unannotated *D. melanogaster* ortholog.

Discussion

Prior to multiplex screening of clones, random sequencing of the whole male reproductive tract library produced 93% (129/139, Table 1.1) accessory gland-derived ESTs, most corresponding to just a few *Acps*. This *Acp* bias cannot be explained by size differences between accessory gland and testis tissues. Instead, *D. mojavensis* testes visually appear much larger than accessory glands (personal observation). Thus, per unit

of tissue, accessory glands must produce many more mRNA transcripts than the testis to account for our observed *Acp* bias. Sequence analysis of our *D. mojavensis* testis EST library generated mostly singleton ESTs (Table 1.2), indicating that the *D. mojavensis* testis transcriptome is much more complex than the accessory gland transcriptome. This observation is consistent with the findings of recent *melanogaster* subgroup male reproductive tract EST projects (Andrews et al. 2000; Swanson et al. 2001).

Previous work has shown that male reproduction genes evolve rapidly relative to other types of genes (Vacquier 1998; Swanson and Vacquier 2002). This pattern is especially striking for *melanogaster* subgroup *Acps* (Begun et al. 2000; Swanson et al. 2001; Kern, Jones, and Begun 2004). BLAST analyses of our *D. mojavensis* male reproductive tract ESTs provided the opportunity to contrast levels of *Acp* vs. testis-expressed gene conservation relative to *D. melanogaster* sequences. Given that both of these classes correspond to genes of male reproduction, it is somewhat surprising that *Acps* are so dramatically less conserved than testis-expressed genes. It would be interesting to know how our testis-expressed genes compare to genes that are more evenly expressed across tissues. For the six genes we analyze that qualify as evenly expressed, the trend is that they are more conserved than testis-expressed genes. However, strong conclusions are not warranted for this small amount of data. Of the 41 total genes from our quantitative PCR analyses that return significant BLAST matches to *D. melanogaster* sequences, only two, *Tes14* and *Tes118*, correspond to putative unannotated genes. Overall, this supports the observation that the *D. melanogaster* genome annotation is of high quality (Drysdale 2003).

The large differences we observed for *Acp* vs. testis-expressed gene conservation highlight the need to recognize distinctions between types of male reproduction genes and the associated implications for adaptive evolution. A key functional difference between *Acps* and testis-expressed genes is that *Acps* are considered to be agents of chemical communication between the sexes (Wolfner 1997) and, therefore, likely targets of postcopulatory sexual selection (Birkhead and Pizzari 2002). These data reinforce this view. To advance our knowledge of this perceived role of *Acps*, future research must determine the female proteins with which they interact. An early step in this direction comes from microarray analyses that identified candidate genes that undergo expression changes in females following mating (Lawniczak and Begun 2004; McGraw et al. 2004). The eventual identification of interacting pairs of male and female postcopulatory proteins will help decipher the specific roles *Acps* play in intersexual chemical communication.

Our quantitative PCR analyses show that most of our surveyed *D. mojavensis* genes are specifically associated with male reproduction. Only six of the 58 genes are roughly equally expressed across tissues. Most of our *D. melanogaster-D. mojavensis Acp* comparisons involve gene pairs with low support for shared ancestry. The *Acp1-2-25* and *Acp53-* family comparison is different. The *D. melanogaster Acp53-* family only includes four known tandemly duplicated genes (Holloway and Begun 2004) with no detectable protein domains. This same family has seven known tandem duplicates in *D. pseudoobscura* (see Chapter 4 for details). Though we cannot be certain how extensive the *Acp1-2-25* family is in *D. mojavensis*, our *D. melanogaster-D. pseudoobscura*

comparative genomics analysis (Chapter 4) leads us to believe the numbers should be similar. Given the apparently small size of this gene family, one might expect that function, and by extension gene expression profiles, would be similar both within and between species. An intriguing finding is that *D. melanogaster Acp53Ea* expression is very similar to *D. mojavensis Acp25* and that *Acp25*/*Acp53Ea* expression patterns are significantly different than *Acp1*/*Acp2* expression patterns ($P < 0.05$). It would be interesting to know if any of the remaining *D. melanogaster Acp53-* genes show increased *Acp* tissue-specificity with expression profiles closer to those of *D. mojavensis Acp1* and *Acp2*.

Expression patterns for putative *D. melanogaster-D. mojavensis* orthologous pairs are generally conserved. All six evenly expressed *D. mojavensis moj-* genes are also evenly expressed in *D. melanogaster*. The 18 testis-expressed *D. mojavensis* genes with putative orthologs maintain testis-expression in *D. melanogaster*. However, our finding that several orthologous pairs differ significantly in degree of testis-specific expression shows that gene regulation has evolved between lineages. This observation is consistent with analyses of gene expression in the *melanogaster* subgroup. Evaluation of genome-wide gene regulation has demonstrated changes in expression between *melanogaster* subgroup species (Meiklejohn et al. 2003; Ranz et al. 2003; Rifkin, Kim, and White 2003). Regulatory changes are particularly evident in male-biased genes, even to the point of showing variation between conspecific populations (Meiklejohn et al. 2003). Given this rapid evolution of *melanogaster* subgroup male-biased gene expression, perhaps the most surprising observation from our comparative data is that all 18 of our

testis-expressed genes are also testis-expressed in *D. melanogaster*. This observation suggests that *Drosophila* genes may often be selected for up- or down-regulation in various tissues, but that changes in tissue-specificity are less common. However, the apparent conservation of tissue-specificity could instead be an artifact of testis-expressed gene sampling bias. With the exception of *Tes100/Tes115*, we only compare expression profiles of orthologs with conserved BLAST similarity, an indication of conserved function. If conservation of gene regulation is positively correlated with coding sequence similarity, we may instead find that changes in tissue-specific expression between orthologous genes are more common. Comparative genomic analyses will identify unalignable orthologous genes to help address this unanswered question.

Chapter 2: Molecular Population Genetics of *Drosophila arizonae* and *D. mojavensis* Male Reproductive Genes

Introduction

Molecular studies in a diverse array of taxa suggest that genes involved in reproduction evolve at an accelerated rate relative to other genes (reviewed in Swanson and Vacquier 2002). Typically, positive selection is implicated as the force driving these changes (Swanson and Vacquier 1995; Metz and Palumbi 1996; Sutton and Wilkinson 1997; Wyckoff, Wang, and Wu 2000; Torgerson, Kulathinal, and Singh 2002; Sorhannus 2003), though the data are insufficient to make generalizations on the relative importance of directional selection vs. genetic drift in these proteins compared to other protein classes. In any case, rapid divergence of reproduction-related proteins, as well as certain functional analyses (see below), are consistent with the notion that male-male and male-female postcopulatory interactions may be associated with rapid divergence between populations and the evolution of reproductive isolation (Eberhard 1996; Rice 1998).

Molecular evolutionary investigation of *Drosophila* reproduction has focused on male accessory gland protein genes (*Acps*) in *D. melanogaster* and its sibling species, *D. simulans*. The number of putative *Acps* in these species is on the order of 83 (Swanson et al. 2001), of which fewer than 20 have strong experimental support (Schafer 1986; DiBenedetto et al. 1987; Monsma and Wolfner 1988; Chen et al. 1988; Wolfner et al. 1997). Certain biochemical functions, including proteases, protease inhibitors, C-type

lectin binding, and lipases appear to be overrepresented in *D. melanogaster*/*D. simulans* *Acps* (Swanson et al. 2001). Genetic analysis has shown that *Acps* are responsible for proper sperm storage (Neubaum and Wolfner 1999; Tram and Wolfner 1999; Chapman et al. 2000), normal ovulation and oviposition (Herndon and Wolfner 1995; Heifetz et al. 2000), and increasing egg-laying rates while reducing female receptivity (Chen et al. 1988; Aigaki et al. 1991; Kalb, DiBenedetto, and Wolfner 1993). *Acps* show much higher rates of protein divergence (Aguadé 1997, 1998, 1999; Tsaur and Wu 1997; Tsaur, Ting, and Wu 1998; Begun et al. 2000; Swanson et al. 2001) and protein polymorphism (Coulthart and Singh 1988) compared to "average" proteins in *D. melanogaster* and *D. simulans* (e.g., Begun et al. 2000). Less energy has been devoted to investigation of genes primarily expressed in testes. However, some anecdotal evidence suggests these genes may also tend to evolve quickly and be associated with evolution of novel function (Long and Langley 1993; Nurminsky et al. 1998; Betrán and Long 2003).

Our current population genetic understanding of *Drosophila* is dominated by data from *melanogaster* subgroup species. Thus, we have no way of knowing whether the patterns of polymorphism and divergence or the functional biology of reproduction-related proteins will be similar in other *Drosophila* species. Given the hypothesis that the dynamics of male-reproduction related proteins may be driven by male-male and male-female postcopulatory interactions, one strategy for furthering our understanding of these proteins is to focus on *Drosophila* species having different reproductive biology from *D. melanogaster* and *D. simulans*. Desert *Drosophila* of the *repleta* group, *D. arizonae* and *D. mojavensis*, share a mating system that differs in many ways from *melanogaster*

subgroup species. Desert *Drosophila* remate much quicker and more often (Markow 2002), males reach reproductive maturity much later (Pitnick, Markow, and Spicer 1995), mating triggers an insemination reaction mass in the female reproductive tract (Patterson and Stone 1952; Markow and Ankney 1988), and much higher proportions of male ejaculates are incorporated into female somatic tissue (Markow and Ankney 1984; Pitnick, Spicer, and Markow 1997). The analyses of reproductive tract genes in *D. arizonae* and *D. mojavensis* reported here aim to address the generality of the evolutionary genetic studies on reproductive genes within the *melanogaster* subgroup and to provide a molecular framework for the ongoing research of reproductive character evolution within the *repleta* group.

Materials and Methods

Isolation and Characterization of *D. mojavensis* Genes

The *D. mojavensis* male reproductive tract and testis EST libraries were the sources from which genes were selected for population genetic analysis (see Chapter 1 for details on library construction). Genes represented by multiple ESTs were preferentially selected under the assumption that they were more likely to be specific to either accessory glands or testes. ESTs for which open reading frames (ORFs) could not confidently be determined were eliminated from consideration (see Chapter 1 for criteria used to identify ORFs). Likewise, if PCR amplification of genomic DNA proved difficult after several attempts with different sets of primers, those ESTs were eliminated

from consideration (with several genes well under 1kb, this outcome was not so uncommon).

Prior to quantitative PCR analysis of the genes, dot blots were used to insure that the genes selected for population genetic analysis included several that were likely enriched for accessory gland or testis expression. Replicate nylon filters containing blots of PCR product from the selected genes were separately hybridized to testis and accessory gland-derived cDNAs to determine tissue specificity. The hybridization temperature was set to 65^{o}C in a buffer consisting of 0.5M NaPi (pH 7.2), 7% SDS, 1mM EDTA. The filter washes were done at 60^{o}C with buffer at 40mM NaPi, 1% SDS, and 1mM EDTA. After the set of genes to be used for population genetic analysis was defined, quantitative PCR was used to more accurately characterize the transcription patterns of all genes surveyed and provided the basis for separating the genes into accessory gland (*Acp*-), testis-expressed (*Tes*-), and all tissue (*moj*-) classes (see Chapter 1 for more detail).

Since relatively few nucleotides were surveyed for most genes (e.g., most *Acps* are small), we expect to have limited power to reject the null hypothesis for individual loci. Therefore, many of our analyses will test hypotheses on groups of genes (e.g., *Acps* vs. testis-expressed genes). The few cases of detailed investigation of individual genes will be presented in a separate section. Most analyses contrast patterns of variation in *Acps* vs. testis-expressed genes. Our criteria for establishing *Acp* and testis-expressed genes (based on $2^{-\Delta\Delta C_T}$ scores greater than five as the threshold—see the quantitative PCR section in Chapter 1 for details) yielded 19 accessory gland genes, 33 testis-

expressed genes, and six genes that are more evenly expressed across tissues. In principle, these six genes can be used to compare male-reproduction-related genes vs. other genes. However, the relatively small sample size means such inferences may be weak.

Population Genetic and Molecular Evolution Analyses

A total of 15 fly stocks from the *Drosophila* Species Stock Center (Tucson, AZ) were used for collection of population genetic data. *Drosophila arizonae* (15081-1271.00, 15081-1271.04, 15081-1271.05, 15081-1271.08, 15081-1271.12, 15081-1271.13, 15081-1271.14; various locations, mainland Mexico) and *D. mojavensis* were represented by seven lines each, while a single *D. mulleri* stock (15081-1371.00; Lake Travis, Texas) provided an outgroup. Of the seven *D. mojavensis* stocks, four were *D. mojavensis baja* (15081-1351.03, 15081-1351.09, 15081-1351.12, 15081-1351.14; various locations, Baja, Mexico) and three were *D. mojavensis mojavensis* (15081-1352.00, 15081-1352.01, 15081-1352.02; various locations, southern California, USA). Primers used for amplification of genomic DNA were designed from library ESTs. The Expand High-fidelity polymerase system (Roche Molecular Biochemicals) was used for PCR amplification. In order to isolate single alleles for sequencing, PCR products were directly cloned into the TOPO vector (Invitrogen, USA) and used for bacterial transformations according to manufacturer's guidelines. Amplified colony PCR products and their associated sequences were obtained using M13 reverse and T7 primers. All

sequencing was done on an Applied Biosystems 377 automated sequencer (ABI, USA). Sequences were aligned and edited using the DNASTAR software package (Lasergene, Madison, WI). The DnaSP program (Rozas and Rozas 1999) was used for most of the population genetic analyses. Average levels of polymorphism or divergence for different groupings of genes refer to weighted means, according to sequence length. For genes sampled for multiple alleles, replacement and synonymous divergence represent the average pairwise difference. Fixations for polarized McDonald-Kreitman tests were assigned using parsimony. Only codons with single mutations that could be clearly assigned to either the *D. arizonae* or *D. mojavensis* lineage were considered.

Lineage-specific estimates of synonymous and replacement divergence were estimated as two free parameters by maximum likelihood with the PAML computer program (Yang 1997). For most of these analyses we used one randomly selected allele from each of three species: *D. arizonae*, *D. mojavensis* and *D. mulleri*. In a few cases for which the *D. mulleri* outgroup was not available, we used a recent duplicated gene predating the *D. arizonae/D. mojavensis* speciation event. The duplicated genes provide a satisfactory outgroup alternative to *D. mulleri* since their synonymous divergence, in all cases, is either comparable or considerably less than average *D. mulleri* synonymous divergence (Table 2.4; see Chapter 3 for additional details). Hypothesis testing was carried out using likelihood ratio tests (Goldman and Yang 1994, Yang 1998). To determine whether or not Ka significantly exceeds Ks in a particular lineage, the likelihood value for the null hypothesis (Ka = Ks) was also calculated. Twice the log

likelihood difference between the two models is then compared to a X^2 distribution with one degree of freedom to determine the level of significance.

Results

We surveyed a total of 56 genes for our population genetics analysis, including estimates of polymorphism and divergence (see Table 2.2). Up to seven lines each of *D. arizonae* and *D. mojavensis* were analyzed for several genes. Most of the remaining genes were characterized by a single allele each from *D. arizonae* and *D. mojavensis*. A single *D. mulleri* allele was sequenced whenever possible for use as an outgroup. Overall, an average of 9.29 alleles and 376 bp were sequenced for each of the 56 genes in this study.

Evidence of *D. m. baja*-*D. m. mojavensis* Population Substructure

Our *D. mojavensis* data consists of up to four alleles of *D. m. baja* and three alleles of *D. m. mojavensis* from various locations of Baja, Mexico and southern California, respectively. Table 2.1 shows our analysis of population substructure between *D. m. baja* and *D. m. mojavensis*. We use the fixation index, F_{ST}, to estimate genetic differentiation between subspecies. Several individual genes show signs of differentiation. *Acp7* notably shows the most evidence of substructure at 0.864. However, because of the very small size of most surveyed genes, we can be more

confident of average F_{ST} values, weighted according to sequence length. The average for all genes is 0.150, with the *Acp* subset of genes slightly higher at 0.168. These results suggest that there are restrictions to gene flow between subspecies, although our average values are within an observed range for different *D. melanogaster* populations (Caracristi and Schlötterer 2003). The *D. melanogaster* survey covered populations from Africa, America, and Europe and estimates an F_{ST} range of 0.004-0.205, with $F_{ST} = 0.205$ corresponding to differentiation between a European and African population.

We also investigate genetic differentiation by estimating divergence between subspecies (Ka and Ks) and comparing those values to nucleotide diversity (π) within subspecies (Table 2.1). Since both measurements represent the probability that a particular nucleotide site drawn from two individuals is different, they can be directly compared. Again, our analysis shows some evidence of population substructure. Averaged across all genes, Ka (0.006) is higher than both replacement *D. m. baja* (0.005) and *D. m. mojavensis* (0.004) nucleotide diversity. However, there are no significant differences between sets of Ka vs. replacement π measurements (Mann-Whitney *U*-test, $P = 0.77$ and $P = 0.41$ for *D. m. baja* and *D. m. mojavensis*, respectively). Any differentiation at synonymous sites is less pronounced with *D. m. baja* synonymous π at 0.016, Ks at 0.015, and *D. m. mojavensis* synonymous π at 0.013.

Because average *D. m. baja*-*D. m. mojavensis* F_{ST} falls within an observed *D. melanogaster* range and between subspecies divergence is not significantly different from within subspecies measurements of nucleotide diversity, we do not distinguish between

D. m. baja and *D. m. mojavensis* alleles in our population genetics analyses below. We expect our estimates of polymorphism to be slightly inflated because of this population substructure. However, our tests of adaptive evolution compare nucleotide substitution patterns at synonymous vs. replacement sites. Therefore, population substructure will either have no effect on our tests (Ka/Ks ratios) or only decrease the probability of detecting adaptive evolution (McDonald-Kreitman tests, see below).

Levels of Synonymous and Replacement Polymorphism and Divergence

Summary statistics for heterozygosity and divergence for individual genes and for gene categories are presented in Tables 2.2-4. As suggested by previously published molecular population genetics data from these species (e.g., Begun and Whitley 2002; Matzkin and Eanes 2003), they are highly variable (Table 2.2). The average synonymous heterozygosity across all surveyed genes for *D. mojavensis* and *D. arizonae* are 0.0181 and 0.0170, respectively (Table 2.3). Synonymous heterozygosity is marginally lower for *Acps* at 0.0135 and 0.0156 compared to testis-expressed genes at 0.0175 and 0.0170, in *D. arizonae* and *D. mojavensis*, respectively. The more evenly expressed *moj-* genes are more polymorphic at synonymous sites at 0.0292 and 0.0346 in *D. arizonae* and *D. mojavensis*, respectively. Since the polymorphism data from the *moj-* class derives from only two genes (*moj9* and *moj30*, Table 2.2), we are unable to conclude that their higher variability reflects a general pattern. Synonymous divergence between *D. arizonae* and *D. mojavensis* is similar across these three gene categories as well (Table 2.3, but see the

polarized analysis below for between species differences). Testis-expressed genes are the most divergent at 0.0682, followed by *Acps* at 0.0643 and *moj-* genes at 0.0518.

The patterns observed for replacement variation are quite different. First, mean replacement heterozygosity of *Acps* in both species is greater than that of testis-expressed or *moj-* genes (Table 2.3). This pattern is especially striking between testis-expressed and accessory gland genes of *D. mojavensis*, with *Acps* about 3.7 times more variable than testis genes in *D. mojavensis* compared to 1.8 times more variable than testis-expressed genes in *D. arizonae*. As expected given the aforementioned patterns, *D. mojavensis Acps* have the highest ratio of replacement to synonymous heterozygosity (0.5991), followed by *D. arizonae Acps* at 0.4866 (Table 2.3). Both are considerably higher than the ratios for testis-expressed genes (*D. arizonae*, 0.2095; *D. mojavensis*, 0.1476) and *moj-* genes (*D. arizonae*, 0.1553; *D. mojavensis*, 0.1308).

Average *Acp* replacement divergence between *D. arizonae* and *D. mojavensis* is also considerably higher (0.0595) than that observed at testis-expressed (0.0128) or *moj-* genes (0.0060). The ratio of replacement to synonymous divergence for *Acps* (0.9257) is 4.9 times greater than the corresponding *Tes-* genes ratio (0.1873), as expected based on the observation of increased accessory gland protein divergence relative to testis-expressed genes, contrasted to similar synonymous divergence for both classes.

A survey of *Acp* variation in *D. simulans* and *D. melanogaster* also suggested that these genes evolve unusually quickly at replacement sites relative to other genes (Begun et al. 2000). However, the relative amount of replacement to synonymous variation at *Acps* in *D. arizonae* and *D. mojavensis* is much greater than that observed in *D. simulans*

and *D. melanogaster*. For example the ratio of replacement to synonymous polymorphism for desert *Drosophila* (0.5991 for *D. mojavensis*, 0.4866 for *D. arizonae*; Table 2.3) is about 2-fold greater than the corresponding ratio in *D. simulans* (0.2643) The same is true for replacement to synonymous divergence, as the Ka/Ks ratio for desert *Drosophila* (0.9257) is more than 2-fold greater than the Ka/Ks ratio for *D. melanogaster/D. simulans* (0.4248). Thus, levels of both protein polymorphism and divergence are considerably greater at *Acps* in *D. arizonae*/*D. mojavensis* than in *D. melanogaster*/*D. simulans*.

The divergence estimates from Tables 2.2-3 result from pairwise comparisons and thus provide no insight into evolution along the *D. arizonae* vs. *D. mojavensis* lineage. We investigated evolution along these two lineages using both parsimony and likelihood-based approaches. Table 2.4 shows the results for all genes for which an outgroup sequence was available. As one might expect from previous analyses, the rank order of Ka/Ks ratios is *Acp* > *Tes-* > *moj-* in each of the three lineages. Moreover, in each lineage there are multiple *Acps* with Ka exceeding Ks. Six genes, all *Acps*, have Ka/Ks > 1 in the pairwise analysis (Table 2.2), while eight of the nine *Acps* have Ka/Ks > 1 in at least one of the three lineages of the polarized analysis (Table 2.4). Several other non-polarized *Acps* have unusually high Ka/Ks values (i.e., greater than 0.5). In contrast, the highest Ka/Ks ratio among non-polarized *Tes-* and *moj-* genes is 0.8992 for *Tes109*, with most genes considerably lower (i.e., less than 0.5). The polarized *Tes-* and *moj-* genes are similarly conserved, with just two examples of Ka/Ks > 1, *Tes105* along the *D. mojavensis* lineage and *Tes114* along the *D. mulleri* lineage. In both cases, however, Ka

estimates are lower than the testis averages for the respective species, and the Ka/Ks > 1 results are largely due to negligible Ks divergence (zero in both cases). Overall, the Ka/Ks ratios for polarized *Acps* and *Tes*- genes are highly significantly different (Mann-Whitney *U*-test, $P << 0.01$).

Though *Acp* proteins are evolving more quickly than *Tes*- or *moj*- genes in desert *Drosophila*, it appears that the relative rates of *Acp* protein evolution vary across lineages. Specifically, the *D. mojavensis* lineage has a considerably greater average *Acp* Ka/Ks ratio than either the *D. arizonae* or *D. mulleri* lineage. Across all nine *Acps*, the Ka/Ks ratio for *D. mojavensis* (2.0776) is 2.4 times greater than the ratio for *D. arizonae* (0.8715). Although *Acp* replacement divergence is higher in *D. mojavensis* (0.0273) than *D. arizonae* (0.0220), the much lower Ks in *D. mojavensis* vs. *D. arizonae Acps* makes a major contribution to the higher *D. mojavensis Acp* Ka/Ks ratio. One possible reason for the low *D. mojavensis* Ks relative to *D. arizonae* Ks could be different patterns of evolution at synonymous sites between lineages. However, our estimates of effective number of codons (Wright 1990) show no major differences between lineages. The averages for *D. mojavensis Acps* and testis-expressed genes, weighted according to size, are 51.8 and 50.8, respectively. The corresponding values for *D. arizonae* are 50.7 and 51.6, respectively. Thus, codon bias of *D. mojavensis Acps* is slightly lower than *D. arizonae Acps*, contrary to expectations if purifying selection at synonymous sites were contributing to the lower *D. mojavensis* Ks values.

Unfortunately, we have *D. mulleri* data from only five *Acps*. This limits our ability to directly compare *Acp* Ka/Ks across the three lineages in a comparable set of

analyses. For these five genes the Ka/Ks average ratio is similar for *D. arizonae* and *D. mulleri* (0.8273 and 0.8484, respectively), while the *D. mojavensis* Ka/Ks ratio (1.7163) is roughly 2-fold greater. Note that the *D. mulleri* data are potentially biased because genes that are evolving more quickly would tend to be underrepresented as a result of PCR failure using primers designed from *D. mojavensis* sequence.

Two *Acps*, *Acp7* and *Acp16a*, have Ka/Ks significantly greater than one in the *D. mojavensis* lineage, while neither gene is significant in *D. arizonae*. The significant Ka/Ks for *D. mojavensis Acp7* reflects a contribution from low synonymous divergence (0.0000), as replacement divergence is similar in *D. mojavensis* (0.0275) to the *Acp* mean (0.0273) for the *D. mojavensis* lineage (Table 2.4). On the other hand, the high Ka/Ks ratio for *D. mojavensis Acp16a* is primarily attributable to the atypically high replacement divergence (0.1538) relative to the lineage *Acp* mean (0.0273). *D. mulleri* provides a solitary example of Ka significantly exceeding Ks with *Acp7* ($P < 0.05$; patterns of evolution at duplicate genes will be discussed in Chapter 3). This is partly due to a higher than average Ka value for *Acp7* (0.2560 compared to 0.1525), though Ks is also lower (0.1200) than the average across all *D. mulleri Acps* (0.1798).

Joint Analysis of Polymorphism and Divergence

According to the neutral theory of molecular evolution, the ratio of replacement to synonymous substitutions should be similar to the ratio of replacement to synonymous polymorphisms (Kimura 1983). The McDonald-Kreitman test uses a 2x2 contingency

table to detect differences in these ratios (McDonald and Kreitman 1991). Table 2.5 shows the polymorphism and fixation data for individual genes at synonymous and replacement sites. For cases in which an outgroup sequence was available (outgroups identical to those in Table 2.4), fixed differences between *D. arizonae* and *D. mojavensis* were polarized using parsimony. Of 54 tests, only one non-polarized (*Acp25*) and two polarized tests (*D. mojavensis*, *Acp1*; and *D. arizonae*, *Acp2*) are significant ($P < 0.05$). Given an adjusted critical value for 54 tests, there is no evidence from analysis of individual genes for deviations from neutrality. However, most of the genes in our survey are small, resulting in limited power for hypothesis testing on a gene-by-gene basis. The lack of power in the single gene analysis motivates the analysis of pooled data (Table 2.6). The 2x2 table for *Acps* is significantly heterogeneous in a direction consistent with adaptive protein evolution, and remains marginally significant if *Acp25* (the single *Acp* that individually deviates from neutrality) is removed from the analysis. Another individual gene that warrants mention is *Acp48*. With a total of 60 mutations to contribute to the 2x2 contingency table, one might speculate that it has a major effect on the overall conclusion. However, removing the *Acp48* data increases the significance of the heterogeneity of the remaining *Acps*. Overall, the analysis of pooled polymorphic and fixed mutations supports the notion that directional selection plays a role in accessory gland protein divergence. Data from testis-expressed and *moj-* genes show no significant deviations from neutral expectation in 2x2 contingency tables.

Further evidence for different evolutionary processes amongst gene-classes can be found in the ratios of replacement fixations to polymorphisms. While a total of seven

Acps have more replacement fixations than polymorphisms, no *Tes*- or *moj*- genes do, with the exception of *Tes112*, which has no replacement polymorphisms and just a single fixation. The ratio of fixed to polymorphic replacement mutations for *Acps* (139:115) is highly significantly different from the ratio for testis-expressed genes (15:60; *G*-test, $P << 0.01$), a result that cannot be easily explained by different neutral mutation rates for the two protein classes. The *moj*- genes ratio (0:16) is more testis-like, though with so few data, strong conclusions are unwarranted.

Investigation of polarized fixations provides more insight into the evolutionary process in the *D. arizonae* and *D. mojavensis* lineages, though at a cost of reduced number of loci and substitutions included in the analysis. The polarized data for different gene classes is presented in Table 2.7. The *Acp* data from *D. mojavensis* show a highly significant ($P = 0.004$) deviation from neutral expectations, most easily interpreted as a large excess of replacement fixations. However, the *D. arizonae Acp* data are not significantly heterogeneous ($P = 0.181$). The results support the idea that directional selection has greater effects on *Acp* divergence in *D. mojavensis* than in *D. arizonae.* The elevated Ka/Ks ratio of *D. mojavensis* vs. *D. arizonae Acps* is consistent with this inference. Note that the numbers of fixed replacement vs. synonymous mutations (24:2) in *D. mojavensis* corresponds to a Ka/Ks ratio for fixed sites of roughly 4, providing additional support for the interpretation that the 2x2 table for *D. mojavensis Acps* can only plausibly be explained by adaptive protein evolution. Polarized data from *moj*- genes in both lineages and testis-expressed genes in *D. mojavensis* are not significantly heterogeneous, while the data from *D. arizonae* are marginally significant (Fisher's exact

test, $P = 0.056$; G-test, $P = 0.026$). In both lineages, however, these non-*Acp* data deviate in the direction of excess replacement polymorphisms rather than fixations, a pattern consistent with purifying selection.

Discussion

Population genetic investigation of accessory gland protein genes has previously focused on *D. melanogaster* and *D. simulans* (Aguadé 1997, 1998, 1999; Tsaur and Wu 1997; Tsaur, Ting, and Wu 1998; Begun et al. 2000; Swanson et al. 2001; Kern, Jones, and Begun 2004). Our study of *Acps* and testis-expressed genes of desert *Drosophila* from the *repleta* group was motivated by our interest in understanding whether the highly diverged mating system of these flies (relative to *D. melanogaster* and *D. simulans*) is associated with different population genetic patterns and mechanisms for male reproduction-related genes.

This question may be especially germane to the issue of *Acps* (rather than testis-expressed genes). Desert *Drosophila* from the *repleta* group remate much more frequently than do *D. melanogaster* or *D. simulans*, opening up the possibility for stronger or fundamentally different selection on male-male and male-female interactions in the *repleta* group. Previous results from within and between species matings of desert *Drosophila* (Patterson and Stone 1952, Knowles and Markow 2001) support the idea of rapid evolution of ejaculate-female interactions. If *Acps* are major players in postcopulatory male-male and male-female interactions (Wolfner 1997, 2002; Chapman

2001), we might expect to observe different patterns of evolution in desert *Drosophila Acps* compared to *melanogaster* subgroup *Acps*. The fact that *D. mojavensis* males make detectable postmating donations to females whereas *D. melanogaster* and *D. simulans* do not (Markow and Ankney 1984; Pitnick, Spicer, and Markow 1997), is another interesting biological difference that could at least in principle affect *Acp* evolution.

Our population genetic analysis of desert *Drosophila Acps* showed some similarities and several important differences with respect to *D. melanogaster*/*D. simulans*. *D. melanogaster* and D. *simulans Acps* are highly polymorphic and divergent at replacement sites compared to "typical" genes in these two species (Begun et al. 2000, Swanson et al. 2001). *Acps* from *D. arizonae* and *D. mojavensis* showed a similar pattern in that they were much more polymorphic and divergent at replacement sites, at least compared to the non-*Acp* genes (mostly testis-expressed genes) surveyed here. However, *D. arizonae*/*D. mojavensis Acps* are proportionally much more polymorphic and divergent in terms of protein variation compared to *D. melanogaster*/*D. simulans Acps* (Table 2.3). One interpretation is that *Acps* tend to be under less functional constraint in desert *Drosophila* compared to the *melanogaster* subgroup. Alternatively, *Acps* could be under stronger directional selection in desert *Drosophila*.

Two types of results support the idea that *Acps* experience directional selection in desert *Drosophila*. First, the Ka/Ks ratio is significantly greater than one for two of nine *D. mojavensis Acps*. Given the small number of bases surveyed per gene and the fact that the Ka/Ks test for positive selection is extremely conservative, observing two of nine genes as individually significant is remarkable. The mean Ka/Ks for *D. mojavensis Acps*

is 2.078, an extremely high value for any class of genes. Second, the McDonald-Kreitman tests provide strong evidence for adaptive protein evolution in *Acps*, but not testis-expressed genes. Interestingly, the *Acp* data strongly deviate from neutral expectations in *D. mojavensis*, but not in *D. arizonae*. Overall, both rates of evolution and contrasts of polymorphic and fixed mutations support the inference of directional selection on accessory gland proteins in the *D. mojavensis* lineage.

Table 2.5 suggests that the highly significant result from the pooled data presented in Table 2.7 is from a consistent excess of replacement fixations across most *D. mojavensis Acps*. This pattern differs from that observed in *D. melanogaster* and *D. simulans Acp* variation, in which a highly significant McDonald-Kreitman test resulting from analysis of 13 *Acps* was attributable mostly to two genes, *Acp26Aa* and *Acp36DE* (Begun et al. 2000). Note that polarized analyses of polymorphic and fixed, synonymous and replacement variation have not been carried out for the *D. melanogaster*/*D. simulans* comparison, as outgroup data are lacking. In this respect, the population genetic inferences for desert *Drosophila* are more incisive than those for *D. melanogaster* and *D. simulans*.

An interesting observation regarding *D. melanogaster* and *D. simulans Acp* polymorphism was that *D. simulans* was proportionally more variable for amino acids (relative to synonymous variants) than was *D. melanogaster* (Begun et al. 2000). This is in contrast to analyses from most genes in these two species suggesting that *D. melanogaster* is proportionally more polymorphic than *D. simulans* at the protein level (Aquadro, Lado, and Noon 1988; Begun 1996). Our analysis of *D. mojavensis* vs. *D.*

arizonae Acp polymorphism revealed no such heterogeneity (Table 2.7; *G*-test, $P = 0.574$), further supporting the notion that the mechanisms of *Acp* protein divergence differ between *D. arizonae* and *D. mojavensis*.

There has been much speculation regarding the potential importance of adaptive protein evolution for male-reproduction related genes. However, the data presented here are the first molecular population genetic analysis of a sample of *Drosophila* genes expressed primarily in testes. Our results show that testis-expressed genes evolve much more slowly than *Acps* and show no evidence of adaptive protein divergence. Thus, at least based on these limited data, we would not conclude that genes associated with male reproduction in *Drosophila* evolve at similar rates or by similar mechanisms. Clearly, however, the functional categorization of genes as testis-expressed vs. *Acp* is somewhat crude. For example, criteria of biochemical function or other attributes of gene function associated with reproduction could reveal a significant role for adaptive protein evolution in many testis-expressed genes. Even so, a degree of generalization is in order. It is well documented that spermatogenesis requires a large set of genes (Fuller 1993; Poccia 1994; Eddy 1998), whose functions are, therefore, unlikely to extend to male-male and male-female postcopulatory interactions. Thus, an important distinction between our set of *Acps* vs. testis-expressed genes is that *Acps* are likely to contain a much larger proportion of genes involved in postcopulatory male-male and male-female interactions (Wolfner 1997, 2002; Chapman 2001). Our data show that proteins likely to be involved in these types of interactions may be common targets for directional selection.

Given their very close evolutionary relationship and similar mating systems, the inference of directional selection on *D. mojavensis Acps* and the lack of such an inference for *D. arizonae* is surprising. One notable distinction between mating systems is that the *D. mojavensis* ejaculate donation to female somatic tissues is almost 10-fold higher than in *D. arizonae*, a difference that is not even remotely matched by any other sister species pairs from a large phylogenetic survey (Pitnick, Spicer, and Markow 1997). Perhaps larger somatic donations reflect an increased *Acp* role in postcopulatory male-female interactions. Data from other species pairs with differences in ejaculate donation must be gathered to determine the role this aspect of *Drosophila* mating systems plays in *Acp* evolution. We note that differences between *D. arizonae* and *D. mojavensis Acp* protein evolution rates do not diminish our inference that the distinctions between *melanogaster* subgroup vs. desert *Drosophila* mating systems explain the different patterns of *Acp* evolution in these groups. Though evidence of adaptive *Acp* evolution is less dramatic for *D. arizonae* than *D. mojavensis*, Ka/Ks is still more than twice as high for *D. arizonae Acps* (0.8715) compared to *D. melanogaster*/*D. simulans Acps* (0.4248).

An alternative explanation of the differences between *D. arizonae* and *D. mojavensis Acp* protein evolution is that our sampling of *Acp* loci has compromised our ability to make an unbiased comparison between lineages. Because our *Acps* were isolated from a *D. mojavensis* accessory gland cDNA library, we are biased toward isolating genes that are more abundantly expressed in *D. mojavensis* than *D. arizonae*. Therefore, a possible explanation for the differential importance of adaptive protein evolution in *D. arizonae* vs. *D. mojavensis* is that more abundantly expressed *Acps* are

under stronger directional selection. This possibility is easily addressed through additional quantitative analysis (for both expression and population genetics) of larger numbers of *Acps* in both species and could help determine the contributing roles of gene regulation and protein change to adaptive evolution.

Chapter 3: Molecular Population Genetics of *Drosophila arizonae* and *D. mojavensis* Accessory Gland Protein Gene Families

Introduction

Postcopulatory conflict between males, in the form of sperm competition, can be an important component of male fitness in polyandrous species (Birkhead and Møller 1998). Numerous strategies have evolved to increase sperm competitive ability, often mediated by components of the seminal fluid (Birkhead and Møller 1998; Chapman 2001; Fry and Wilkinson 2004). Females also have an interest in paternity and can play an important role in deciding the outcome of sperm competition (Eberhard 1996; Birkhead and Pizzari 2002; Bernasconi et al. 2004). Thus, postcopulatory sexual selection drives male adaptations to increase sperm competitive ability and female counter-adaptations to bias paternity, maintaining a state of antagonistic coevolution between the sexes (Rice 1996, 1998). Consistent with this hypothesis, proteins that mediate fertilization are known to evolve rapidly in many species (Vacquier 1998; Swanson and Vacquier 2002). Accordingly, postcopulatory interactions and the molecules behind them have drawn considerable attention for their potential role in generating reproductive isolation between populations (Parker and Partridge 1998; Rice 1998; Pitnick, Markow, and Spicer 1999; Arnqvist et al. 2000; Gavrilets 2000; Knowles and Markow 2001).

In *Drosophila*, empirical studies suggest that there is abundant genetic variation affecting traits related to male-male and male-female postcopulatory interactions. For example, *D. melanogaster* males that were allowed to evolve to a genetically static female environment caused a decrease in female survivorship, increased remating rates, and increased seminal fluid toxicity when mated back to these same female flies (Rice 1996). Moreover, *D. melanogaster* male- and female-expressed variation significantly affects patterns of sperm use in multiply mated female flies (Clark et al. 1995; Clark and Begun 1998).

Male accessory gland proteins (*Acps*) of the *melanogaster* subgroup have received most of the attention as potential molecular agents of male-male and male-female postcopulatory interactions in *Drosophila*. *Acps* are a major component of *Drosophila* seminal fluid. There are an estimated 83 *Acps* in the *melanogaster* subgroup (Swanson et al. 2001). *Acps* have been shown to stimulate ovulation and increase egg laying rates (Kalb, DiBenedetto, and Wolfner 1993; Herndon and Wolfner 1995; Heifetz et al. 2000), bind sperm and effect sperm storage (Neubaum and Wolfner 1999; Tram and Wolfner 1999), effect the outcome of sperm competition (Harshman and Prout 1994; Chapman et al. 2000), decrease female receptivity (Chen et al. 1988; Aigaki et al. 1991; Chapman et al. 2003; Liu and Kubli 2003), and decrease female life span (Chapman, Hutchings and Partridge 1993; Chapman et al. 1995; Lung et al. 2002). One experiment found a correlation between sperm displacement phenotypes and SSCP haplotypes at some *Acp* loci (Clark et al. 1995). *Acps* evolve rapidly in the *melanogaster* subgroup (Begun et al. 2000, Swanson et al. 2001; Kern et al. 2004), in at least some cases as a result of

directional selection (Tsaur and Wu 1997; Aguadé 1998; Tsaur, Ting, and Wu 1998; Aguadé 1999; Begun et al. 2000; Kern, Jones, and Begun 2004).

The genus *Drosophila* is highly diverse and includes taxa with mating systems that differ dramatically from *melanogaster* subgroup flies (Powell 1997). Desert *Drosophila* of the *repleta* group, *D. arizonae* and *D. mojavensis*, are a case in point. For example, males of desert *Drosophila* reach reproductive maturity at 4-5 days post-eclosion, compared to two days for *D. melanogaster* (Pitnick, Markow, and Spicer 1995). Male age at reproductive maturity is positively correlated with sperm size and the size of female sperm-storage organ in *Drosophila* species (Pitnick, Markow, and Spicer 1995, 1999). Moreover, sperm size and sperm-storage organ size are coevolving rapidly in *D. mojavensis*, with geographically distinct populations expressing divergent phenotypes of these correlated traits (Pitnick et al. 2003). Another difference between these taxa is female remating, which occurs much more rapidly and more often in desert *Drosophila* (Markow 2002). Within 24 hours of an initial mating, roughly 95% of *D. mojavensis* females remate (Markow 1982). In contrast, only 2% of *D. melanogaster* females remate in this same time period (Pyle and Gromko 1981). Higher remating rates in desert *Drosophila* could potentially increase selection on phenotypes related to postcopulatory male-male or male-female interactions (Markow 2002; Singh, Singh, and Hoenigsberg 2002).

Additional differences between *repleta* group and *melanogaster* subgroup flies are evident in the short-term physiological response of females following copulation. Transfer of seminal fluid triggers an insemination reaction within the female reproductive

tract of desert *Drosophila* (Patterson and Stone 1952) but is diminutive in *D. melanogaster* (Wheeler 1947; Markow and Ankney 1988). This insemination reaction, which is superficially similar to inflammation, results in a mass in the female reproductive tract. Remating does not occur during the several hours that it persists (Patterson 1947; Knowles and Markow 2001). The intensity of the insemination reaction is highly variable, with interspecific matings (e.g., *D. arizonae* and *D. mojavensis*) triggering an exaggerated and harder mass, which persists significantly longer than within species insemination reactions (Patterson 1947). Interestingly, exaggerated insemination reactions are observed in some crosses between geographically distinct populations of *D. mojavensis*, suggesting that interpopulation postcopulatory incompatibilities may evolve very quickly (Knowles and Markow 2001). Finally, ejaculate components of many *repleta* group species, including *D. mojavensis*, are incorporated into female somatic tissues, a phenomenon not known to occur in the *melanogaster* subgroup (Markow and Ankney 1984; Pitnick, Spicer, and Markow 1997).

There are many differences between desert *Drosophila* and *melanogaster* subgroup flies in postcopulatory phenotypes that are likely to be mediated by *Acps*. Our earlier results suggested that although general patterns of protein variation in *Acps* from desert *Drosophila* and *melanogaster* subgroup flies are similar, there are important quantitative differences between groups. For example, we found faster rates of protein evolution and stronger evidence for directional selection in *repleta* group *Acp* comparisons relative to *melanogaster* subgroup comparisons (Chapter 2). Previous analysis of 13 annotated *D. melanogaster Acps* suggested that recent *Acp* gene

duplications are rare in the *D. melanogaster*//*D. simulans* lineage (Holloway and Begun 2004). Here we report the discovery of several recent *Acp* duplications in *D. arizonae* /*D. mojavensis*. Our analyses suggest that several of these recent duplications have diverged under directional selection, a phenomenon not observed in *D. melanogaster* (Holloway and Begun 2004). These data provide additional support for different evolutionary processes acting on *Acps* in these lineages, perhaps as a result of mating system divergence.

Materials and Methods

Fly stocks, PCR amplification, and sequencing methods are the same as in Chapter 2. *Acps* identified from the original *D. mojavensis* reproductive tract library ESTs (see Chapter 1) bear the suffix "-*a*" while subsequently identified duplicates follow alphabetically according to their order of discovery.

Gene Discovery

Duplicate *Acps* described here were accidentally amplified as secondary PCR products from primers designed from *D. mojavensis* accessory gland ESTs. Thus, these duplicate *Acps* are not a random sample of potential *D. mojavensis Acp* duplications. Rather, they should be biased towards relatively low sequence divergence. Sequence data from each putative duplicate *Acp* were used to design PCR primers for amplifying additional copies of each duplicate for population genetic analysis. However, the very

short length of some *Acps* under investigation made it difficult to isolate duplicates from all of the fly lines used in our earlier analysis of *Acps* (Chapter 2).

Organization of Duplicated *Acps*

Patterns of sequence divergence (see below) in most cases provided unambiguous evidence that the *Acps* in question are duplications rather than highly diverged alleles. Nevertheless, two types of analyses were used to further investigate the duplication hypothesis and provide insight into genomic organization of duplications.

Under the premise that recent duplications are often tandemly arranged, we designed PCR primers from putative duplicates to amplify across intergenic regions. We used LA-Taq long PCR polymerase (TaKaRa, Japan) with an extension time of ten minutes and cycling parameters according to manufacturers instructions. Successfully amplified fragments were end-sequenced to confirm that the amplified product corresponded to the expected genomic sequence under the tandem duplication hypothesis. Second, we used data from the NCBI *D. mojavensis* WGS trace archive (2,038,648 sequences in the database; August, 2004) to construct partial assemblies of putative duplicate *Acp* regions. We performed BLASTn analysis of putative duplicate *Acps* to the trace archive, followed by additional BLAST based sequence walks. The resulting traces were assembled in the SeqMan program of the DNASTAR software package (Lasergene, Madison, WI).

Ka/Ks Estimation and Hypothesis Tests of Adaptive Protein Evolution

Nucleotide distances were used to infer the topologies of duplicate family genealogies. Maximum-likelihood estimation of branch-specific Ka and Ks values used the free-ratio model of the PAML computer program (Yang 1997). For inferences that required outgroups, outgroups were determined by pairwise distance estimates and corroborated by PAML branch length output. For genes sampled for multiple alleles, one random allele was chosen for PAML analyses. Alignments were generated using the DNASTAR software package (Lasergene, Madison, WI), and manually adjusted where appropriate. Indel variation for codon positions that were gapped in > 50% of the aligned sequences were omitted from the analyses. To test whether the Ka value of a given branch significantly exceeds the Ks value, the likelihood ratio test was used to compare the free-ratios model to the null model (Ka/Ks = 1). Twice the log-likelihood difference was then compared to a X^2 distribution with one degree of freedom to determine significance levels.

Results

Evidence of Gene Duplication

In the course of our molecular population genetic analysis of 18 *Acps* in *D. arizonae* and *D. mojavensis* (Chapter 2), sequence data from four genes revealed alleles that were unusually highly diverged from the majority of alleles sampled. These genes

were clearly related to the target genes, but had levels of divergence that in most cases could only be plausibly interpreted as evidence of gene duplication (Table 3.3). Under the assumption that duplicate *Acps* likely originated through tandem duplication, we designed PCR primers from the putative duplicates to amplify intergenic sequence between paralogs. We were able to amplify intergenic sequences for *Acp5a-c*, *Acp16a-b*, *Acp21a-b*, and *Acp27a-b*, thereby confirming their duplicate status. Thus, there is at least one experimentally confirmed tandemly duplicated pair in each of the four *Acp* groups discussed here. Table 3.1 provides a summary of fly lines that have been verified by PCR to carry particular putative duplicate gene copies.

Of the putative duplications not verified by PCR to be tandemly arranged, all except *D. mojavensis Acp5a* and *Acp5b* have synonymous divergence levels (Table 3.3) that easily surpass levels of variation consistent with heterozygosity at a locus in *Drosophila* (Moriyama and Powell 1996; Begun and Whitley 2002). However, other data support the view that *Acp5a* and *Acp5b* are true duplicates rather than highly diverged alleles. For example, high synonymous *Acp5a vs. Acp5b* divergence within *D. arizonae* strongly supports the paralogy hypothesis, as the value far exceeds typical levels of heterozygosity (Table 3.3). This suggests that the low level *Acp5a* vs. *Acp5b* Ks within *D. mojavensis* might be due to stochasticity associated with the small number of silent sites surveyed (n = 24 alignable synonymous sites, Table 3.3), though gene conversion is also a possible explanation. However, the very high *D. mojavensis Acp5a* vs. *Acp5b* Ka also supports paralogy. Finally, a distance tree for *Acp5-* duplicates clusters the *D.*

arizonae and *D. mojavensis Acp5b* alleles. Overall, there is little doubt our *Acp5a* and *Acp5b* alleles represent duplicated genes in both *D. arizonae* and *D. mojavensis*.

Physical Organization of Duplications

Our PCR and NCBI *D. mojavensis* trace archive assembly data provide evidence on the genomic organization of related duplicates. *Acp5a* is approximately 5 kb 5' of *Acp5c*; the genomic location of *Acp5b* is uncertain. Our partial assembly of NCBI *D. mojavensis* trace archive sequences reveals ~2 kb of distinct 5' and 3' flanking DNA for *Acp5b* and *Acp5c*. Therefore, if *Acp5b* is tandemly arranged, it is likely 4 kb or more 3' of *Acp5c* or 2 kb or more 5' of *Acp5a*.

PCR data for *Acp16*- duplicates show that *Acp16b* is approximately 3 kb 5'of *Acp16a*. Trace archive assembly of this duplicate region reveals *Acp16b* to be 2.8 kb upstream of the *Acp16c* region. Unfortunately, *Acp16c* coding region was not found in the trace archive. Therefore, we cannot confirm the location of *Acp16c*, though our PCR data taken together with the trace archive data suggest *Acp16a* and *Acp16c* are probably less than 500 bp apart. PCR data for *Acp21* duplications and *Acp27* duplications clearly demonstrate tandem organization for both, with *Acp21b* approximately 1.5 kb 5' of *Acp21a* and *Acp27a* approximately 3 kb 5' of *Acp27b*.

Polymorphism and Interspecific Divergence of Duplicate *Acps*

Polymorphism and interspecific orthologous divergence of duplicate *Acps* is presented in Table 3.2, along with weighted averages of the statistics and comparable data from putative single copy *D. mojavensis* and *D. arizonae Acps*. Though there is considerable variation among duplicate genes for synonymous and replacement polymorphism, the small number of sites surveyed per gene precludes any speculation about heterogeneous forces. Overall, silent heterozygosity is slightly lower in duplicated *D. arizonae* and *D. mojavensis Acps* compared to single copy *Acps* from these species. Silent divergence between species is also slightly slower for duplicated vs. single copy *Acps*. In contrast, replacement heterozygosity and divergence are higher for duplicated *Acps* than for single copy *Acps* in *D. arizonae* and *D. mojavensis*, with replacement divergence marginally significantly higher for duplicate *Acps* (Chapter 2, Mann-Whitney one-tailed *U*-test only, $P = 0.0457$). The average Ka/Ks ratio for duplicated *Acps* is 2.12 (Table 3.2), which is significantly higher than the ratio for single-copy *Acps* from these species (Chapter 2, Mann-Whitney one-tailed *U*-test, $P = 0.00679$; one value left out of each group because Ks = 0) and higher than the ratio for *Acps* in *D. melanogaster* vs. *D. simulans* comparisons (Swanson *et al*. 2001). *Acp21a*, with an orthologous Ka/Ks ratio of 4.12 stands out as a particularly rapidly evolving protein.

Paralogous Ka/Ks Ratios

Paralogous divergence estimates of duplicate *Acps* are given in Table 3.3. The ratio of replacement to synonymous divergence exceeds one for each pairwise comparison. Paralogous Ka/Ks estimates exceed four in at least one pairwise comparison for three of the four *Acp* families. In the other group, *Acp16*-, the highest value is for *D. arizonae Acp16a-b* (1.934). These extremely high Ka/Ks estimates seem particularly noteworthy given that only one of the 14 putative single-copy *Acps* investigated in *D. arizonae*/*D. mojavensis* has an interspecific Ka/Ks ratio that is comparable to the estimates we report here (*Acp119*, Ka/Ks can not be calculated because Ks = 0; Chapter 2). Of the remaining 13 putative single-copy *Acps*, the highest ratio is 1.374, corresponding to *Acp1*.

Dating Duplications Relative to *D. arizonae*/*D. mojavensis* Speciation

We were able to sequence duplicated genes from both species for several loci, clearly indicating that duplication occurred prior to speciation. However, four duplicates are only documented in one of the species: *Acp5c*, *Acp16c*, *Acp21b*, *Acp27b*. In these cases we use divergence estimates to roughly calibrate the time of gene duplication relative to speciation. Note, however, that the small size and rapid protein divergence (often under selection, see below) of duplicate genes can complicate the task of dating duplications relative to speciation.

Synonymous and replacement divergence of paralogous pairs for both *Acp5c* and *Acp16c* (*c* vs. *a* and *c* vs. *b*, Table 3.3) exceed those same estimates for interspecific divergence of *Acp5*- and *Acp16*- (*ari -a* vs. *moj -a* and *ari -b* vs. *moj -b*, Table 3.2). Thus, we conclude that these two duplications occurred prior to speciation. Under this hypothesis, our inability to sample *D. arizonae* alleles might be explained by interspecific divergence that was too great to amplify *D. arizonae* alleles using primers designed from *D. mojavensis* DNA sequences.

Paralogous vs. orthologous divergence for *Acp21a-b* displays the opposite pattern. For both synonymous and replacement divergence, the *D. arizonae Acp21a-b* paralog measurements are lower than the corresponding *Acp21a* ortholog divergence measurements (Tables 3.2-3). Thus, we conclude that the *D. arizonae Acp21b* gene arose by duplication subsequent to speciation.

The *D. mojavensis Acp27b* duplicate is more difficult to date relative to *D. arizonae*/*D. mojavensis* speciation. Both synonymous and replacement measurements for paralog divergence are greater than the corresponding *Acp27a* ortholog measurements (Tables 3.2-3). However, the differences between the two, especially with respect to synonymous divergence, are too small to support a strong conclusion. The replacement divergence is 10X greater between paralogs than between orthologs (0.134 compared to 0.013). However, this could be due to adaptive evolution of *Acp27b* (see below). Synonymous divergence may provide a more reliable estimate of the age of the duplications. Here the differences are not as dramatic, with *Acp27a* ortholog synonymous divergence at 0.006 and *Acp27a-b* paralog synonymous divergence at 0.021.

This supports duplication prior to speciation. Note, however, that this *Acp27a* interspecific Ks value (0.006) is very low relative to the average interspecific synonymous divergence for all duplicate *Acps* (0.044, Table 3.2). With the *Acp27a-b* paralog synonymous divergence at an intermediate value relative to interspecific synonymous divergence of *Acp27a* and all duplicate *Acps*, we conclude that gene duplication occurred either before or close to the time of speciation.

Our maximum-likelihood estimates of branch distance lengths from the analyses below are in agreement with all of the above conclusions. Furthermore, the estimated *D. mojavensis Acp27b* branch distance is more than 7X greater than either the *D. arizonae* or *D. mojavensis Acp27a* branches (0.245 vs. 0.032 and 0.010, respectively). These figures, however, are largely driven by replacement rather than synonymous substitutions.

Branch-Specific Divergence of Duplicate *Acps*

Maximum-likelihood analysis of the *Acp5* duplicate gene family reveals very high rates of protein evolution along most gene-tree branches (Fig. 4.1); only the *D. arizonae Acp5a* branch has Ka/Ks < 1. The *D. mojavensis Acp5c* branch has a Ka/Ks ratio that is significantly greater than one ($P < 0.05$). However, this inference was from an unrooted tree. Thus, we cannot determine how much of this accelerated protein evolution occurred on the branch leading to *Acp5c* vs. the branch leading to the common ancestor of *Acp5a*/*Acp5b*.

Evolution of the *Acp16* gene family also reveals very high rates of protein evolution, with five of the six branches for which Ka/Ks could be estimated having Ka/Ks > 1 (Fig. 4.2). The *Acp16a* branch prior to speciation shows the strongest evidence of adaptive evolution with Ka/Ks = 13.6, which is significantly greater than one ($P < 0.05$). Though the *D. mojavensis Acp16a* branch and the *Acp16b* pre-speciation branch both have Ka/Ks values greater than two, neither is significantly greater than one.

Branch-specific divergence estimates for *Acp21* and *Acp27* are shown in Table 3.4. Both gene families generally show little synonymous divergence and very high levels of replacement divergence. Overall, four of the six branches have Ka/Ks significantly greater than 1. Note that of the 18 branches for which Ka/Ks was estimated in PAML (Fig 4.1, Fig 4.2, and Table 3.4), the estimate was greater than one for 15 branches. Overall, the data provide extremely strong support for the hypothesis that directional selection has played a prominent role in protein evolution of duplicated *Acps* in *D. arizonae*/*D. mojavensis*.

McDonald-Kreitman Tests of Adaptive Evolution

Joint analysis of polymorphism and divergence can be used to test for adaptive evolution. According to the neutral model, the ratio of replacement to synonymous polymorphic mutations should be the same as the ratio of replacement to synonymous fixations (McDonald and Kreitman 1991). A McDonald and Kreitman (MK) test uses a 2x2 contingency table to test for deviations from the neutral model. In order to sample

all branches from a duplicate gene tree for our MK tests, we sampled some branches more than once. Thus, duplicate pairings are not necessarily independent of one another. Table 3.5 presents our MK tests for duplicate *Acps*, including polarized analyses whenever possible. There are no examples of significant heterogeneity for any of the tests. This is not entirely surprising, given the small size of these genes. However, the distribution of polymorphism is somewhat unusual with one pairing having more synonymous polymorphism, 20 pairings more replacement polymorphism, and the rest either no or equal synonymous vs. replacement polymorphism. Despite a lack of total independence between pairings, the pattern of excess replacement polymorphism is very clear. This is consistent with pooled MK tests of a larger sample of *D. arizonae*/*D. mojavensis Acps* displaying greater numbers of replacement polymorphisms (115 vs. 63, Chapter 2). In contrast, polymorphism data from 13 pooled *D. melanogaster Acp* genes shows 158 synonymous polymorphisms and 142 replacement polymorphisms (Begun et al. 2001). We cannot be certain as to why there is proportionally more replacement polymorphism in desert *Drosophila* than *D. melanogaster*. However, it is possible that natural selection is elevating replacement polymorphism in desert *Drosophila*. With this skew in distribution of polymorphisms, MK tests are much less effective in proving adaptive divergence of genes.

Acp27- presents a noteworthy MK result. Divergence between *D. mojavensis Acp27a-b* paralogs is borderline significant ($P = 0.0504$, Fisher's exact test; Table 3.5). There are a total of 17 replacement fixations, compared to zero synonymous fixations. Polarization of these data using *D. arizonae Acp27a* shows that all 17 mutations occurred

along the *Acp27b* branch. Given that we predict either a star phylogeny or *Acp27b* duplication prior to speciation (see above), it is not technically correct to use *D. arizonae Acp27a* as an outgroup. Nevertheless, we can say that none of these 17 replacement fixations are attributable to the *Acp27a* gene after speciation. Instead, the rapid evolution has either occurred along the *Acp27a* branch before speciation, the *Acp27b* branch, or some combination of the two.

Discussion

Sequence analysis of *Acp* genes from the *melanogaster* subgroup has demonstrated that this class of seminal fluid proteins evolves rapidly relative to other classes of genes (Begun et al. 2000; Swanson et al. 2001; Kern, Jones, and Begun 2004). This rapid evolution is often interpreted as evidence of natural selection, which is thought to play an important role in sperm competition and male-female postcopulatory interactions (Rice 1996; Swanson and Vacquier 2002). We have previously shown that *Acp* genes of *D. arizonae* and *D. mojavensis* evolve more rapidly than *melanogaster* subgroup *Acps* (Chapter 2), an observation that is consistent with expectations based on their dramatically different mating systems (Markow 1996, 2002). Here we show that four *D. arizonae*/*D. mojavensis Acp* gene families evolve more rapidly than putative single-copy *Acps*, with evidence of adaptive evolution in all four families. These results are consistent with observations suggesting gene duplication can facilitate adaptive protein evolution (Ohno 1970; Ohta 1994; Li 1995). Interspecific Ka/Ks ratios for all

duplicate Acps varied from 0.808 to 4.121, significantly exceeding the distribution of Ka/Ks ratios for putative single-copy *Acps*. Moreover, paralogous Ka/Ks ratios were even higher, demonstrating a broad timeframe for adaptive evolution since most of these duplicates predate *D. arizonae*/*D. mojavensis* speciation. Our maximum-likelihood analyses show that 17 out of 20 duplicate gene tree branches have Ka/Ks ratios greater than one. At least one branch from each duplicate gene family significantly exceeds Ka/Ks = 1.

Another apparent difference between desert *Drosophila* and *melanogaster* subgroup *Acps* is in the history of duplicate gene fixation events. There are three known duplicate *Acp* families in *D. melanogaster*: the *Acp53Ea* family (Holloway and Begun 2004), *Acp29AB*/*Lectin29Ca*/*Lectin30A* (Holloway and Begun 2004), and *Acp70A*/*Dup99B* (though *Dup99B* is expressed in the male ejaculatory duct; Saudan et al. 2002). These paralogs are not readily alignable at the nucleotide level and are, therefore, considerably older than the *D. arizonae*/*D. mojavensis* duplications we discuss here. Furthermore, there is no reason to believe that we have identified all of the recent *D. arizonae*/*D. mojavensis Acp* duplications. Theory suggests that the probability of new duplicate gene fixation increases with positive selection for functional divergence, or neofunctionalization (Walsh 1995; Lynch et al. 2001). The compelling evidence in support of adaptive protein divergence in these four desert *Drosophila Acp* families is consistent with neofunctionalization. The disparity between duplication histories of desert *Drosophila* and *melanogaster* subgroup *Acp* families could be a result of stronger directional selection for neofunctionalization in *D. arizonae*/*D. mojavensis*.

A potential side effect of gene duplication and neofunctionalization is the generation of reproductive isolation through chromosomal repatterning (Lynch and Force 2000; Lynch et al. 2001). For example, if a duplication event involving a wild type allele rises to a high frequency in a population, followed by neofunctionalization of the ancestral locus, a map change between populations develops. Postzygotic reproductive isolation arises passively because of the possibility of obtaining hybrid offspring that produce null gametes for the ancestral function of such loci. *D. mojavensis* presents an ideal model system to test this model of speciation. A recent study demonstrated geographical variation for hybrid male sterility in *D. mojavensis* (Reed and Markow 2004). Population genetic evidence demonstrating duplicate *Acp* map changes between these populations would support this hypothesis, especially if correlations to hybrid fitness are revealed. We note that the relative chromosomal positions of several duplicate *Acps* discussed here are currently unknown. The assembly of the *D. mojavensis* genome will enable the comparison of duplicate *Acp* map positions between populations.

Polymorphism data also reveals interesting differences between desert *Drosophila* duplicate *Acps* and putative single copy *Acps*. Duplicate *Acps* are about 54% and 58% less polymorphic at synonymous sites but about 143% and 185% more polymorphic at replacement sites vs. single copy *Acps* in *D. arizonae* and *D. mojavensis*, respectively (Table 3.2). Taking the absolute number of polymorphic sites for duplicate vs. single copy *Acps* (Table 3.5 and Table 2.5, Chapter 2), there is significant synonymous vs. replacement site heterogeneity both within *D. arizonae* and *D. mojavensis* (*G*-test, $P = 0.014$ in both cases). Given the high rate of adaptive evolution at these loci, and the

evidence for significant geographical variation in postcopulatory *D. mojavensis* phenotypes (Knowles and Markow 2001; Pitnick et al. 2003; Reed and Markow 2004), some of the replacement polymorphism in *Acp* gene families might be due to divergent selection between geographically isolated populations. Additional population genetics data comparing intra- and interpopulation dynamics between conspecific desert *Drosophila* populations are needed to resolve this question.

Chapter 4: Comparative Genomics of Accessory Gland Protein Genes In *Drosophila melanogaster* and *D. pseudoobscura*

Introduction

Much of comparative genomics research seeks to detect putative functional elements (e.g., genes) by virtue of sequence conservation (e.g., Batzoglou et al. 2000; Wiehe et al. 2001; Jaillon et al. 2003). However, from the evolutionary perspective, rapidly evolving genes are as interesting as slowly evolving genes because genes experiencing directional selection are more likely to be rapidly evolving. An understanding of biological diversity and adaptation will require evolutionary and functional analysis of rapidly evolving genes. The gain or loss of genes over time must also be explained. For example, microorganisms that take on an obligate intracellular lifestyle often lose genes (e.g., Moran 2003). Over long time periods, even conserved proteins can be lost in certain lineages (Krylov et al. 2003, Kortschak et al. 2003). Nonetheless, our general understanding of gene loss is likely plagued by ascertainment bias. For example, genes that are prone to loss over relatively shorter time scales may tend to evolve quickly and therefore more likely to be unannotated in model system genomes. Gain and loss of genes is intriguing because it suggests the possibility that "homologous" functions can be partially (or even mostly) coded for by non-homologous proteins. The population genetic mechanisms of gene loss are also interesting. For example, gene loss could represent decay of a "non-essential" gene under mutation

pressure, a change of the biology in a lineage that renders a previously essential gene dispensable, or removal of a gene by selection (Olson 1999, Galvani and Slatkin 2003, Olson and Varki 2003). We would like to distinguish among these possibilities.

Drosophila is an attractive model system for addressing these questions. Flies have relatively compact genomes for animals, and the deep annotation and experimental tractability of the model fly, *D. melanogaster*, provide an excellent starting point for investigating the functional and evolutionary biology of rapidly evolving proteins. The only *Drosophila* species other than *D. melanogaster* with a "complete" genome sequence is *D. pseudoobscura* (the *D. pseudoobscura* genome sequence is estimated to have 7X euchromatic coverage, corresponding to approximately 99.9% of the euchromatic genome; BGM-HGSC, http://www.hgsc.bcm.tmc.edu/projects/drosophila/), though other species are currently being sequenced. *D. pseudoobscura* diverged from the *melanogaster* group approximately 21-46 million years ago (Beckenbach et al. 1993). An initial comparison of these species covering more than 300 kb of sequence data shows that the majority of *D. melanogaster* Release 3 gene models are highly conserved in *D. pseudoobscura*, and that microsynteny is largely maintained in *D. pseudoobscura* and *D. melanogaster* (Bergman et al. 2002).

Data from animals suggest that the portion of the genome coding for reproduction-related function may be unusually dynamic. For example, an interesting generality emerging from studies of molecular evolution is the relatively rapid evolution of proteins associated with male reproduction (e.g., Swanson and Vacquier 2002). In *Drosophila*, testis and accessory gland proteins (*Acps*) show rapid divergence (Coulthart

and Singh 1988; Begun et al. 2000; Swanson et al. 2001; Kern, Jones, and Begun 2004) compared to other proteins. Three known genes contributing to reproductive isolation in flies (Ting et al. 1998, Barbash et al. 2003, Presgraves et al. 2003) evolve extremely quickly, suggesting that rapidly evolving genes may play an important role in speciation. Anecdotal evidence is consistent with the notion that reproduction-related *Drosophila* proteins may be gained or lost unusually frequently (e.g., Long and Langley 1993; Nurminsky et al. 1998; Betrán and Long 2003).

Drosophila Acps have probably received more population genetic attention than any other class of reproduction-related gene in flies. Males transfer *Acps* to females during mating. They have been implicated in induction of oviposition, in rendering females recalcitrant to re-mating, and in mediating sperm displacement and sperm storage in females (Neubaum and Wolfner 1999; Tram and Wolfner 1999; reviewed in Wolfner 2002; Heifetz and Wolfner 2004). As noted previously, *Acps* evolve quickly compared to other *Drosophila* proteins. Some of this rapid evolution is likely the result of directional selection (Tsaur, Ting, and Wu 1998; Aguadé 1998; Begun et al. 2000; Holloway and Begun 2004), though the fraction of *Acp* proteins under positive selection is still unclear.

These previous observations of *Drosophila* molecular evolution motivate the work reported here, which addresses three main questions regarding molecular evolution and gain/loss of *Acps* in the *D. melanogaster* vs. *D. pseudoobscura* comparison. First, how does one identify orthologous, rapidly evolving genes that may be sufficiently diverged so as to preclude identification through simple BLAST comparisons between genomes. Second, what are the patterns of protein evolution for highly diverged genes.

Third, and perhaps most interesting, to what extent are rapidly evolving proteins likely to be lineage-restricted – that is, absent in at least some lineages. This last question is especially interesting to us because gene presence/absence variation could be an important aspect of the unique biology of particular lineages, and reproduction-related genes may be more likely than other types of genes to show lineage-restricted distributions. Here we use computational and molecular approaches to investigate these questions using comparison of 13 annotated *Acp* genes from the *D. melanogaster* reference sequence to the *D. pseudoobscura* genome sequence.

Materials and Methods

Computational Analysis

The *D. pseudoobscura* genome (August 2003, Freeze 1 Assembly; BGM-HGSC, http://www.hgsc.bcm.tmc.edu/projects/drosophila/) was screened through extensive BLAST analysis (version 2.2.9; Altschul et al. 1997) for the presence of 13 *D. melanogaster Acps*. A combination of BLAST methods was used to investigate presence/absence of *D. pseudoobscura* orthologs. tBLASTn (peptide sequence query to all six possible reading frames of a nucleotide database) searches of all *D. melanogaster Acps* were performed. *D. melanogaster Acp* flanking sequence was also analyzed in order to establish larger scale homology and microsynteny (or lack thereof) between species. Depending on the immediate genomic neighborhood of individual *Acps*, this

either involved tBLASTn analysis of flanking genes, BLASTn (nucleotide to nucleotide query) analysis of non-coding intergenic sequence, or some combination.

The search for homologous *D. pseudoobscura* sequence began with tBLASTn analysis of *D. melanogaster Acps*. We used E < e-4 as our significance threshold. All potential *D. pseudoobscura* ortholog candidates were BLASTp analyzed back to *D. melanogaster* predicted proteins. To eliminate non-orthologous genes with shared domains or from gene families, only candidates that hit the original *D. melanogaster Acp* at the lowest E score were considered further (there were no ambiguous cases in which a *D. melanogaster Acp* E score was close to the score from another gene). Proximal and distal flanking sequence was then analyzed for all 13 *Acps*. Starting from immediate flanking sequence and moving out in both directions, non-coding intergenic sequence and neighboring genes were BLAST analyzed. Flanking sequences were typically queried in 2-4 kb intervals but exact lengths depended on the genetic neighborhood of individual *Acps*. Flanking genes were analyzed in the same manner as the *Acps* described above. The same E score threshold (E < e-4) was used for intergenic sequence BLASTn analysis, but additional hits (E < 0.05) to *D. pseudoobscura* microsyntenic sequence were also noted once homology was already established. For every *D. melanogaster Acp*, the amount of flanking sequence analyzed was dictated based on certainty of homology. For example, if 2 kb of flanking sequence produced five intergenic BLASTn hits of E < e-10 each, we did not necessarily analyze additional sequence from that flank.

D. pseudoobscura Acp ortholog candidate regions, as defined by microsyntenic patterns within homologous genomic segments, were further analyzed for the presence of

open reading frames (ORFs) and evidence of transcription. Computational analysis of *D. pseudoobscura Acp* ortholog candidate regions consisted of identifying potential ORFs that showed similarity to *D. melanogaster* counterparts in amino acid similarity, ORF length, intron/exon structure, protein domains, or presence/absence of putative signal sequences. The SignalP 3.0 server (hidden Markov method) was used to detect putative signal peptides (Nielsen and Krogh 1998, Bendtsen et al. 2004). SignalP probabilities range from 0 to 1, with 1 indicating very high probability of a signal peptide. NCBI CD-Search was used to identify conserved domains (Marchler-Bauer et al. 2003). Protein sequences were aligned using the default Clustal parameters of MegAlign in the DNASTAR software package (Lasergene, Madison, WI). Protein similarity was calculated as the number of identical residues/total number of alignable residues.

Empirical Methods

Two approaches, RACE and reverse Northerns, were used to empirically investigate transcription in *D. pseudoobscura* genomic regions that are homologous to regions containing *Acps* in *D. melanogaster*.

RACE templates were separately produced from sexually mature male and female *D. pseudoobscura* flies from a stock that combined two isofemale lines originally collected by M. Noor. mRNA from each sex was isolated using the MicroPolyA-Pure kit (Ambion, Austin, TX). RACE-ready cDNA was prepared and target molecules were PCR amplified and isolated using the GeneRacer (Invitrogen) kit according to the

manufacturers instructions. The protocol separates the truncated from the complete and mature mRNA products, preferentially selecting the full-length transcripts for first-strand cDNA synthesis. Target-specific primers were paired with either 3' or 5' RACE primers to amplify candidate transcripts. In many cases, multiple target primers were used. RACE was performed on pooled aliquots of male and female RACE-ready cDNA. Amplified products were cloned into the TOPO vector (Invitrogen, USA) and used for bacterial transformations according to manufacturer's instructions. Direct sequencing of colony PCR products was carried out on an Applied Biosystems 3700 sequencer (ABI, USA).

Though RACE should be sensitive to low transcript abundance, failure of RACE to amplify a transcript could be a result of suboptimal gene-specific primers. This is a particular concern for small putative transcripts, where primer design options can be limited. Therefore, regions providing no evidence of transcription from RACE reactions were subjected to reverse Northern analysis. Unlike RACE, this approach has the virtue of requiring no specific inferences regarding details of putative protein-coding regions. Candidate and control *D. pseudoobscura* genomic regions were PCR-amplified (all were 4kb or shorter in length). Roughly 500ng of each product were electrophoresed through each of two replicate 1.0% agarose gels and transferred to nylon filters. Separate male and female cDNA probes were prepared from RACE-ready cDNA by ^{32}P-labeling using the Prime-It II kit (Stratagene). These probes were hybridized overnight to the replicate filters at 65^{o}C in a buffer consisting of 0.5M NaPi (pH 7.2), 7% SDS, 1mM EDTA. Filters were washed at 60^{o}C in 40mM NaPi, 1% SDS, 1mM EDTA. The resulting

membranes were exposed to X-ray film to infer evidence of transcription in males and female *D. pseudoobscura*.

Population Genetics

Isofemale lines derived from flies collected by M. Noor were used for population genetics analysis. The Expand High-fidelity polymerase system (Roche Molecular Biochemicals) was used for PCR amplification. In order to isolate single alleles for sequencing, PCR products were directly cloned into the TOPO vector (Invitrogen, USA) and used for bacterial transformations according to manufacturer's guidelines. Amplified colony PCR products and their associated sequences were obtained using M13 reverse and T7 primers. All sequencing was done on an Applied Biosystems 3700 sequencer (ABI, USA). Sequences were assembled and edited using the SeqMan program of the DNASTAR software package (Lasergene, Madison, WI). Summary statistics and the McDonald-Kreitman test of neutral molecular evolution (McDonald and Kreitman 1991) were computed using DnaSP version 3.53 (Rozas and Rozas 1999).

Results

Putative regions of *D. melanogaster*/*D. pseudoobscura* homology based on conserved microsynteny of BLAST matches around individual *D. melanogaster Acps* are depicted in Figs. 4.1-4.12. *D. melanogaster* chromosomal regions are oriented with the 5' end of the *Acps* to the left of the figure. To the right, chromosomal regions are labeled

"proximal' or "distal" to orient the sequences with respect to centromeres and telomeres. Genes are represented by open rectangles, with no breaks for introns except for a few cases in which higher resolution is necessary. Solid horizontal arrows depict the 5' to 3' orientation of genes. Dashed arrows between *D. melanogaster*/*D. pseudoobscura* chromosomal segments are used to depict homologous sequence as determined by BLAST analysis. Dotted horizontal lines indicate intergenic sequence that produced significant BLASTn results. Results from reverse Northerns used to detect *D. pseudoobscura* transcription in candidate *Acp*-containing regions are shown in Fig 4.13. Table 4.1 provides a summary of putative orthology for all *Acps*. Table 4.2 lists the accession nos. for all *D. pseudoobscura* microsyntenic regions, as well as CDS starting positions for orthologous *Acps*. Detailed results for individual *Acps* are reported below, grouped according to evidence of presence (both in the expected microsyntenic region and elsewhere within the genome) and evidence of absence.

Evidence of Gene Presence

Acp26Aa&Ab

These two *Acps* are tandemly arranged in *D. melanogaster*, with the termination codon of *Acp26Aa* less than 200bp proximal to the initiation codon of *Acp26Ab*. Figure 4.1 shows an illustration of the putative homology between *D. melanogaster* and *D. pseudoobscura* in the *Acp26Aa&Ab* region. *Acp26Aa* showed no BLAST similarity to any *D. pseudoobscura* sequence, while *Acp26Ab* generated only a marginally significant

(E = 0.045) tBLASTn hit. Nevertheless, investigation of nearby flanking sequences revealed strong evidence for a *D. pseudoobscura* region of homology on chromosome 4, the correct arm given the homology of *D. pseudoobscura* 4 and *D. melanogaster* 2L (Lakovaara and Saura 1982; Steinemann, Pinsker and Sperlich 1984).

The first 2 kb immediately proximal to *D. melanogaster Acp26Aa* generates five highly significant and contiguous BLASTn hits averaging 41 bp (from E = 3e-10 to E = 6e-05) to a portion of *D. pseudoobscura* chromosome 4 (region a, Fig. 4.1). The 4.5 kb region immediately distal to *Acp26Ab* was similarly characterized by four highly significant and contiguous BLASTn hits averaging 70 bp (from E = 3e-19 to E = 2e-09) to the same *D. pseudoobscura* chromosome 4 contig (partially depicted by region b, Fig. 4.1). Given the contiguous physical organization of the flanking regions in the two species and given the fact that the marginally significant *Acp26Ab* tBLASTn hit fell within the hypothesized microsyntenic 5.1kb region in *D. pseudoobscura* spanning BLASTn hits a and b (Fig. 4.1), it is highly likely that we had identified the homologous region in *D. pseudoobscura*. If a copy of *D. pseudoobscura Acp26Aa* were present, it too, would likely be within this 5.1 kb region.

RACE analysis of this 5.1 kb *D. pseudoobscura* chromosome 4 sequence was used to identify the putative transcripts corresponding to *Acp26Aa* and *Acp26ab*. One gene specific primer for 5' RACE was designed from sequence corresponding to the *D. pseudoobscura* tBLASTn hit for *Acp26ab*. Six additional 5' RACE primers were designed from the 3 kb of *D. pseudoobscura* sequence immediately upstream of the tBLASTn hit to *Acp26Ab*. The rationale for this was that at least one of these six primers

should amplify a portion of a *D. pseudoobscura Acp26Aa* ortholog if it existed within this homologous region. DNA sequences of the resulting successful RACE reactions on *D. pseudoobscura*-derived mRNA, and comparison of these RACE products to genomic sequence clearly revealed both genes. Both genes have conserved intron/exon structure relative to the orthologous *D. melanogaster* genes (Table 4.1). Moreover, both putative *D. pseudoobscura Acps* have strongly predicted signal peptides, as do their putative *D. melanogaster* counterparts (Table 4.1). Male-specific transcription within the *D. pseudoobscura Acp26Aa* candidate region (Fig. 4.13) provides additional support for orthology. The predicted *D. pseudoobscura Acp26Aa* protein is 250 residues (compared to 264 in *D. melanogaster*). The predicted *D. pseudoobscura Acp26Ab* protein is 90 residues (compared to 92 in *D. melanogaster*). Interestingly, in spite of the compelling evidence for orthology, the predicted proteins are extraordinarily diverged, especially *Acp26Aa*. Predicted protein sequences of *D. pseudoobscura* and *D. melanogaster Acp26Aa* are only 18.5% similar (in other words, essentially unalignable). Predicted *Acp26Ab* protein orthologs are more conserved at 33.3% similarity.

Acp32CD

D. melanogaster Acp32CD is closely flanked by neighboring genes (Fig. 4.2). Less than 800bp of intergenic region separates *Acp32CD* from its proximal neighbor, *CG14913*. On the opposite flank, and on the minus strand, the last exon of *CG31868* is 992bp distal to the *Acp32CD* initiation codon. The remaining exons of *CG31868* are

more than 14 kb away from this last exon. All three of these genes generated clear tBLASTn hits to a single, small contiguous region of *D. pseudoobscura*, chromosome 4. Figure 4.2 shows the preserved microsynteny between *D. melanogaster* and *D. pseudoobscura* in this gene region. Of the three, *CG14913* is the most highly conserved gene (E = 2e-79), followed by the last exon of *CG31868* (E = 1e-27), and *Acp32CD* (E = 9e-12). *D. pseudoobscura Acp32CD*, like its *D. melanogaster* ortholog, is a single exon gene with an apparent signal peptide sequence (Table 4.1). The *D. pseudoobscura Acp32CD* protein contains 299 residues, compared to 252 residues in *D. melanogaster*. The difference in size is largely due to the middle section of the *D. pseudoobscura* protein, which contains a section of several glycine residue repeats. Even so, the orthologs show 43.7% similarity.

Acp53Ea and Duplicates

Acp53Ea is one of four tandemly duplicated genes in *D. melanogaster* found in a region just over 3 kb in length (Fig. 4.3). Two of these genes, *Acp53C14a* and *Acp53C14b*, are proximal to *Acp53Ea*. Paralogous *D. melanogaster* protein divergence is 48.5% between *Acp53Ea* and *Acp53C14a*, 42.5% between *Acp53Ea* and *Acp53C14b*, and 45% between *Acp53C14a* and *Acp53C14b* (Holloway and Begun 2004). The other duplicate, which will be referred to here as *Acp53C14c*, is immediately distal to *Acp53Ea*. *Acp53C14c* was previously unannotated and was discovered as a secondary tBLASTn hit to *Acp53C14b* (E = 5e-06). It is the most diverged of the duplicates, at

>65% divergence relative to the other three. Similar gene structures, predicted protein lengths, and strongly predicted signal peptides for all four genes (Table 4.1) support the hypothesis that they are related through repeated tandem duplication.

tBLASTn comparisons of each of the four duplicates to the *D. pseudoobscura* genome revealed corresponding orthologs on chromosome 3, thereby suggesting that these duplications pre-date the *D. melanogaster/D. pseudoobscura* split (E scores for *Acp53C14c*, *Acp53Ea*, *Acp53C14b*, and *Acp53C14a* are 3e-15, 9e-13, 1e-28, and 4e-26, respectively). *Acp53C14c* was found near the endpoint of one *D. pseudoobscura* chromosome 3 contig, while the other three were located contiguously on another chromosome 3 contig. However, further scrutiny of the *Acp53C14c* contig strongly suggests that *Acp53C14c* is likely just upstream of the other *Acp53*- genes, just as it is in *D. melanogaster*. This inference comes from the observation that in *D. pseudoobscura*, *CG8566* (tBLASTn, E = 0.0) is just under 3kb to the left of *Acp53C14c* (orientation as in Fig 4.3), while in *D. melanogaster CG8566* is about 2.2 kb to the left (distal to) *Acp53C14c*. Protein similarity leaves little doubt as to the true orthology of these duplicates, as the most similar interspecific pairings is consistent with conserved microsynteny between species (40.5%, 41.7%, 48.5% and 55% similarity for *Acp53C14c*, *Acp53Ea*, *Acp53C14b*, and *Acp53C14a*, respectively). Thus, the four *D. melanogaster* duplicates predate the *D. melanogaster/D. pseudoobscura* split, as all four are clearly present in *D. pseudoobscura*.

A major difference between these species in this region is that *D. pseudoobscura* has three additional tandem duplicates (*Acp53C14f*, *Acp53C14e*, and *Acp53C14d*)-,

between *Acp53C14c* and *Acp53Ea* (Fig. 4.3). tBLASTn analysis of the *D. melanogaster Acp53C14b* gene originally identified *Acp53C14d* as a weak match (E = 0.001). Additional tBLASTn analysis of *Acp53C14d* to the *D. pseudoobscura* genome revealed the last two duplicates through E scores of 2e-06 (*Acp53C14f*) and 5e-04 (*Acp53C14e*). None of these additional duplicates appear to have *D. melanogaster* orthologs. tBLASTn analysis of all three back to the *D. melanogaster* genome only produced one significant hit for *Acp53C14d* to *D. melanogaster Acp53C14b* (E = 2e-05), and two non-significant hits for *Acp53C14d* to *D. melanogaster Acp53C14a* (E = 0.13) and *D. melanogaster Acp53Ea* (E = 0.28). Neither *Acp53C14e* nor *Acp53C14f* BLASTs registered even weak hits to *D. melanogaster*. Therefore, these additional *D. pseudoobscura* duplicates either originated in the *D. pseudoobscura* lineage or were lost from the *D. melanogaster* lineage.

Evidence of Gene Presence Associated with Genomic Rearrangement

Acp62F

D. melanogaster Acp62F is an intronless gene that codes for a 115 residue protein with a trypsin inhibitor domain and a predicted signal peptide sequence. The nearest distal gene, *CG32296*, is 11 kb away. *CG1240* is the nearest proximal gene, at about 20 kb away. Nevertheless, BLASTn analysis of 3 kb of intergenic sequence along each genomic flank revealed a microsyntenic region to *D. pseudoobscura* chromosome XR (Fig 4.4). The 5' flank is characterized by five, highly significant BLASTn matches, (from E = 2e-18 to E = 2e-8) which average 52 bp in length (region a, Fig. 4.4). The 3'

flank is similarly characterized by four highly significant BLASTn matches, (from E = 6e-18 to E = 2e-11), which average 54 bp in length (region b, Fig. 4.4).

An *Acp62F* ortholog could not be identified in the *D. pseudoobscura* candidate microsyntenic region (between BLASTn matches a-b in Fig. 4.4). Computational analysis of this 3.4 kb region revealed six candidate ORFs, ranging from 62 to 155 residues in length. None of these candidates showed good evidence of a signal peptide sequence (SignalP probabilities ranged from 0 to 0.35) or a trypsin inhibitor domain. RACE analysis of all six possible candidates also failed to detect any evidence of *D. pseudoobscura* transcription. Finally, a PCR product spanning the complete *D. pseudoobscura* candidate region failed to hybridize to male- and female-derived radio-labeled cDNA (Fig. 4.13).

Despite the lack of evidence for a putative *D. pseudoobscura Acp62F* homolog in the expected *D. pseudoobscura* microsyntenic region, tBLASTn analysis of *D. melanogaster Acp62F* revealed three highly significant ortholog candidates (E = 8e-17, 2e-11, and 4e-10, for candidates 1-3 respectively) at different positions of *D. pseudoobscura* chromosome 3 (not tandemly arranged). All three *D. pseudoobscura* ortholog candidates were then BLASTp analyzed back to *D. melanogaster* predicted proteins. Candidate 3 was eliminated from consideration, as its strongest match was another *D. melanogaster* trypsin inhibitor domain protein, CG5267. The two remaining candidates returned *D. melanogaster Acp62F* at the lowest E score (2e-18 and 1e-13 for candidates 1 and 2, respectively). Both *D. pseudoobscura Acp62F* ortholog candidates hit the *D. melanogaster* chromosome 3L gene *CG33259* secondarily (E = 8e-17 and E = 1e-

11 for candidates 1 and 2, respectively). tBLASTn of *D. melanogaster CG33259* back to *D. pseudoobscura* sequences hits candidates 1 and 2 at the lowest E scores (8e-17 and 2e-11 for candidates 1 and 2, respectively). As is the case for *D. melanogaster Acp62F*, both *D. pseudoobscura* ortholog candidates, as well as *D. melanogaster CG33259*, have predicted signal peptides (*P* = 0.985, 0.955, and 0.999 for candidates 1, 2, and *CG33259*, respectively) and contain trypsin inhibitor domains. Gene organization is also similar to *Acp62F*, as *D. pseudoobscura* candidates 1 and 2, and *D. melanogaster CG33259* are single exon genes (135 and 120, and 119 residues for candidates 1, 2, and *CG33259*, respectively). Intergenic flanking sequence analysis of the *D. pseudoobscura* candidates clearly identified microsyntenic tBLASTn homology (from E = 5e-28 to E = 4e-15 for each of the four flanks) to different portions of *D. melanogaster* chromosome 2R, the correct arm given the homology of *D. melanogaster* 2R and *D. pseudoobscura* chromosome 3 (Steinemann, Pinsker, and Sperlich 1984). In both cases, there were no gene annotations in the corresponding *D. melanogaster* microsyntenic region, nor was there evidence of ORFs containing signal peptide sequences or trypsin inhibitor domains. Thus, there is no evidence that any of these trypsin inhibitor domain genes have orthologs within the appropriate microsyntenic regions.

The tBLASTn evidence suggests *D. pseudoobscura* candidate 1 is most likely orthologous to *D. melanogaster Acp62F* if a true ortholog exists in *D. pseudoobscura*. Our RACE analysis of this putative ortholog proves that it is transcribed and intronless as expected. A protein distance tree puts *D. melanogaster Acp62F* and *CG33259* as the most closely related pair, followed by *D. pseudoobscura* candidate 1 and then *D.*

pseudoobscura candidate 2. Thus, candidate 1 is more similar to *Acp62F* than is candidate 2. Given the possibility that the shared trypsin inhibitor domains obscure the evolutionary relationships as a result of convergent or parallel evolution, we also carried out a distance analysis with the shared domains removed (the domain covers 54-55 residues in all four genes). Though similarities decreased as expected, the structure of the distance tree remained the same. *D. melanogaster Acp62F* and *CG33259* are 51.9% similar across the complete proteins. *D. pseudoobscura* candidate 1 is 41.6% similar to *Acp62F*. All other pair wise comparisons are below 38% similar. With domains removed, *D. melanogaster Acp62F* and *CG33259* are 32.7% similar and *D. pseudoobscura* candidate 1 is 30.9% similar to *Acp62F*. All remaining pair wise comparisons drop below 25%.

We conclude that *D. pseudoobscura* candidate 1 is orthologous to *D. melanogaster Acp62F* and that microsynteny has been disrupted as a result of genomic rearrangement in one or both lineages. Given that the gene is on different Muller elements in the two species, a transposition event is likely. We also propose that *D. melanogaster Acp62F* and *CG33259* are related through a duplication event that occurred subsequent to the *D. melanogaster*/*D. pseudoobscura* split. *D. pseudoobscura* candidate 2 is likely either related through a more ancient duplication (and lost in *D. melanogaster*) or is similar through parallel or convergent evolution. However, the shared trypsin inhibitor domain and lack of microsyntenic conservation between species precludes a definitive assessment of orthology from our data.

Acp70A

D. melanogaster Acp70A protein is 55 amino acids long, with 38.33 residues coded by the first of its two exons. It shows clear evidence of a signal peptide sequence (p = 1.0) but no evidence of a conserved protein domain. tBLASTn analysis of *Acp70A* provided no evidence of a *D. pseudoobscura* ortholog in. However, analysis of 4 kb of the 5' flank and 2 kb of the 3' flank indicated that this portion of map region 70A is homologous to a portion of *D. pseudoobscura* chromosome XR (Fig. 4.5). A total of seven small nucleotide segments returning highly significant BLASTn matches (from E = 4e-35 to E = 9e-7; regions a-d, Fig. 4.5) and averaging 55 bp support the hypothesis of homology. The regions of similarity are contiguous between species, with the exception of a pair that indicate a likely micro-inversion event (Fig. 4.5, region b). Accounting for this apparent micro-inversion, if a *D. pseudoobscura* ortholog were present in this microsyntenic region, it could be on the plus strand between regions b-c, or on the minus strand between regions a-b.

Given a small first exon (115 bp of the ORF), there were approximately nine candidate *D. pseudoobscura* first exons within regions a-c. However, only one of the nine carried the signature of a signal peptide sequence (SignalP, $P = 0.969$). Neither 5' nor 3' RACE reactions attempted using primers designed from this first exon candidate succeeded in amplification of *D. pseudoobscura* cDNA. Furthermore, hybridization of *D. pseudoobscura* cDNA to a PCR fragment spanning region a-c provided no evidence of

a transcribed gene in the region (Fig. 4.13), suggesting a microsyntenic ortholog is unlikely.

The most significant tBLASTn result from comparison of *D. melanogaster Acp70A* to the *D. pseudoobscura* genome was E = 0.002, a value sufficiently large to be ignored in most cases. However, closer scrutiny of this weak tBLASTn hit provided additional information. The hit was to chromosome 4 and was identical at 13 of 14 residues from the second exon. Computational analysis and successful 5' RACE amplification of the corresponding region of *D. pseudoobscura* revealed a potential gene with the same intron/exon structure as *D. melanogaster Acp70A* and a strongly predicted signal peptide (SignalP, *P* = 1.0). The candidate protein is 57 residues, two residues longer that the *D. melanogaster Acp70A* protein, with one additional residue in each of the two *D. pseudoobscura* exons (Table 4.1). BLASTp analysis of the predicted *D. pseudoobscura Acp70A* protein to predicted *D. melanogaster* proteins hit only one, *Acp70A* (E = 2e-05), supporting the hypothesis of orthology. Protein alignment of the putative orthologs shows 54.7% similarity.

Analysis of the flanking regions of the *D. pseudoobscura Acp70A* ortholog suggested that the gene is located in a region homologous to region 35F in *D. melanogaster*, between *CG31819* and *CG12455*. BLASTn analysis of the putative *Acp70A* gene in *D. pseudoobscura*, including 4kb of each genomic flank, generated 13 highly significant and contiguous results to this region, averaging 91bp in length (E scores from E = 5e-56 to E = 8e-7 for five 5' flank matches and eight 3' matches). There is no computational evidence for a microsyntenic *D. melanogaster* gene within the space

between 3' and 5' flank BLASTn hits. In fact, this region comprises 4.6 kb in *D. pseudoobscura*, compared to only 590 bp in *D. melanogaster*. We conclude that both species possess a copy of *Acp70A*, though they are located in non-syntenic locations as a result of genome rearrangement, probably transposition between Muller elements.

Evidence of gene absence

Acp29AB and *lectin-29Ca*

These two genes are highly diverged, tandem duplicates in *D. melanogaster* (Holloway and Begun 2004). tBLASTn analyses of both genes was complicated by the lectin domain they share with many fly genes. tBLASTn analysis of both genes to *D. pseudoobscura* returned several significant results (E < 10e-10 leaves eight *Acp29AB* hits and seven *lectin-29Ca* hits). However, the most significant BLAST results for each of the predicted *D. pseudoobscura* proteins back to *D. melanogaster* predicted proteins were to several lectin domain-containing genes other than *Acp29AB* or *lectin-29Ca*. tBLASTn analysis of three neighboring genes allowed us to identify the region in *D. pseudoobscura* that is homologous to the *D. melanogaster Acp29AB/lectin-29Ca* region (Fig. 4.6). These three genes returned highly significant tBLASTn results (*CG17814*, *CG31893* and *CG13394* returned E scores of 5e-17, 5e-28, and 1e-111, respectively) to a single contiguous region of *D. pseudoobscura* chromosome 4.

The major difference in the organization of the microsyntenic region in the two species is that *D. pseudoobscura* has only 145 bp between the termination codon of

CG31893 and the initiation codon of *CG13394*. The same region in *D. melanogaster* covers 2.2kb and contains both *Acp29AB* and *lectin-29Ca*. These data clearly demonstrate that *Acp29AB* and *lectin-29Ca* cannot reside in the homologous region in the two species. We also found no evidence from tBLASTn analysis for a chromosomal rearrangement, as we observed for *Acp62F* and *Acp70A*. Therefore, we conclude that *Acp29AB* and *lectin-29Ca* could only be present in *D. pseudoobscura* given a model of extreme sequence divergence and genomic rearrangement.

Acp33A

tBLASTn analysis returns no significant hits for either of two potential isoforms of *Acp33A*. The nearest gene, *CG6541*, is almost 5 kb distal to *Acp33A*. BLASTn comparison of 3 kb of 5'-flanking sequence to *D. pseudoobscura* generated no significant results. However, BLASTn comparison of the next 2.5 kb of 5' flanking sequence did return a highly significant result to a *D. pseudoobscura* chromosome 4 contig and consisting of 10 contiguous nucleotide segments, averaging 73 bp each (E scores from E = 4e-31 to E = 3e-10; Fig. 4.7, section a). BLASTn of 2 kb of 3' flanking sequence reveals a second highly significant set (E scores from E = 4e-15 to E = 3e-10; Fig. 4.8, section b) of seven contiguous hits averaging 63 bp in length to the beginning of another *D. pseudoobscura* chromosome 4 contig. If there has been no major evolutionary change in the organization of this region, the two *D. pseudoobscura* contigs would be about 3.5 kb apart. However, our long PCR attempts to span the putative *D. pseudoobscura* genome sequence gap were unsuccessful. Although our evidence provides no support for

an *Acp33A* ortholog in *D. pseudoobscura*, assembly of the homologous *D. pseudoobscura* contigs is necessary before any conclusions can be reached.

Acp36DE

Acp36DE is located in a gene-poor region of the *D. melanogaster* genome. It resides in the large first intron of *CG5803*. *Acp36DE* is 35 kb proximal to the first exon of *CG5803* and 24 kb distal to the second exon. There are no other annotated genes in this 59 kb interval.

tBLASTn comparison of *D. melanogaster Acp36DE* to the *D. pseudoobscura* genome revealed no evidence for a *D. pseudoobscura Acp36DE* homolog. However, BLASTn analysis using 5'- and 3'-flanking *D. melanogaster* sequences revealed clear evidence for a region of microsynteny in the two species. Analysis of 3.5 kb of 5' flanking sequence to *Acp36DE* returned highly significant scores (from E = 2e-30 to 6e-6; Fig. 4.8, region a) that consisted of four small stretches of nucleotide matches, averaging 57 bp in length. Similarly, BLASTn analysis of 1.5 kb of 3'-flanking sequence revealed hits for six small DNA segments averaging 42 bp in length and which had E-values ranging from E = 5e-14 to E = 2e-4 (Fig. 4.8, region b). The highly similar proximal-to-distal linear organizations of these small regions in the two species provide strong evidence of microsynteny.

However, two pieces of evidence suggest that there is no *D. pseudoobscura* ortholog of *Acp36DE*. First, the physical scale of the homologous region in the two species suggests that the size of the *D. pseudoobscura* region is insufficient to harbor

Acp36DE. The *D. melanogaster Acp36DE* CDS covers 2739 bp and includes two exons. The second exon is considerably larger, coding for 843 of the 912 protein residues. Nevertheless, the homologous region of *D. pseudoobscura* spans only 1471 bp (Fig. 4.8). The largest possible ORF (including those not starting with methionine) in this region of *D. pseudoobscura* is less than 1/8 of the length of the *D. melanogaster* second exon (309 bp in *D. pseudoobscura* compared to 2531 bp in *D. melanogaster*). Finally, our molecular data provide no evidence in *D. pseudoobscura* for transcripts in the region corresponding to the *Acp36DE* transcript region of *D. melanogaster* (Fig. 4.13).

Acp63F

D. melanogaster Acp63F is located within the 2.3 kb first intron of *CG1065*, but on the opposite strand. *CG1065* is highly conserved in *D. pseudoobscura*, providing the necessary flanking sequence to establish homology to a small region of *D. pseudoobscura* XR (Fig. 4.9). Exons 2-4 generate significant tBLASTn homology proximal to *Acp63F* (E = 4e-67, 2e-74 and 2e-74 for exons 2-4, respectively). Distally, the small first exon of *CG1065* also generates a highly significant hit to this same region of chromosome XR (E = 2e-14, BLASTn only due to small exon size of 13 residues). tBLASTn analysis of *Acp63F* produced no significant or even marginal hits to the *D. pseudoobscura* genome.

The intron-exon organization of *CG1065* is conserved between the two species. However, there is a major difference between *D. melanogaster* and *D. pseudoobscura* in

the size of the first intron, which defines the boundaries of the *Acp63F* gene region in *D. melanogaster*. The intron is almost five times larger in *D. melanogaster* than in *D. pseudoobscura* (2.3 kb versus 470 bp, respectively). The candidate region which would contain the *D. pseudoobscura Acp63F* ortholog can be further refined by noting that there is a small stretch of apparently conserved intron 1 nucleotides (26/27 identical to *D. melanogaster*) within 61 bp of the *D. pseudoobscura CG1065* first exon. Thus, the *D. pseudoobscura* genomic region that would contain *Acp63F* (start to stop codon) is 383 bp. The *D. melanogaster Acp63F* genomic sequence from start to stop codon (including introns) is 361 bp. Including putative 5' and 3'-flanking UTRs, the *D. melanogaster* region is 432 bp. Therefore, it seems rather unlikely that the *D. pseudoobscura Acp63F* gene would fit within this much smaller piece of DNA. Finally, and most importantly, our molecular experiments provide no evidence for *D. pseudoobscura* transcripts associated with the region that would contain *Acp63F* based on patterns of microsynteny in the two species (Fig. 4.13).

Acp76A

D. melanogaster Acp76A is a relatively large accessory gland gene, consisting of a 994 bp first exon, a 69 bp intron and a 173 bp second exon. The *Acp76A* protein contains a serpin domain. Figure 4.10 illustrates BLAST results comparing the *D. melanogaster Acp76A* gene region to the *D. pseudoobscura* genome sequence. BLASTn analysis of a 2 kb region of 5'-flanking DNA revealed three contiguous matches (E

ranging from 1e-28 to 2e-08) averaging 80 bp. BLASTn comparison of 2 kb of 3'-flanking DNA returned a highly significant result (E ranging from 8e-26 to 2e-10) of five contiguous nucleotide sequences averaging 83 bp each. These regions correspond to *D. pseudoobscura* chromosome XR. The amount of genomic DNA defined by these regions of sequence similarity is about 2.3 kb in *D. melanogaster*, but only 1031 bp in *D. pseudoobscura*. Thus, given the size of the *D. melanogaster* transcript (1235 bp from start to stop, intron included), it seems unlikely that there would be sufficient genomic sequence to harbor a similarly structured *D. pseudoobscura* homolog. Furthermore, this candidate *D. pseudoobscura* region shows no BLAST similarity to *D. melanogaster Acp76A*; its largest possible ORF is only 61 residues or 183 bp, which is considerably shorter than the 994 bp first exon of *D. melanogaster Acp76A*. Finally, we found no evidence of a *D. pseudoobscura* transcript associated with the 1235 bp candidate region of DNA (Fig. 4.13).

Although the microsyntenic region does not appear to contain a *D. pseudoobscura Acp76A* ortholog, we observed two weakly significant tBLASTn hits to *Acp76A* from other parts of the *D. pseudoobscura* genome. The strongest hit was to chromosome 3 (E = 2e-06), but was ruled out as a true ortholog based on the fact that a tBLASTn search of its predicted peptide sequence back to *D. melanogaster* genes returned over 20 serpin-domain containing genes with considerably lower E scores than the *Acp76A* score (E = 3e-9 for *Acp76A*, compared to a low of E = 3e-63 for *CG9456*). The other weakly significant tBLASTn hit to this gene in *D. pseudoobscura* comprised two contiguous stretches of peptide sequence to a non-syntenic portion of chromosome XR (E = 7e-04).

When compared back to *D. melanogaster* predicted proteins, the candidate peptide sequences only returned *Acp76A* as a significant BLASTp hit (E = 7e-7). However, the corresponding *D. pseudoobscura* genomic sequence does not appear to contain a viable candidate ortholog. The putative peptide sequences correspond to residues 199-239 and 271-298, both from the first exon of *D. melanogaster Acp76A*. The similar sequences in *D. pseudoobscura* are in the proper order but are separated by 65 bp, negating the possibility of a single continuous reading frame covering both matches. Moreover, the largest possible ORF that includes either of these putative peptide sequences is only 60 residues, less than 1/5 of the amino acid sequence coded for by the first exon in *D. melanogaster*. Additionally, several attempts to amplify RACE products associated with this candidate sequence failed, suggesting that transcription within this region is unlikely.

Acp95EF

D. melanogaster Acp95EF predicted protein is 52 residues long and has a strongly predicted signal sequence (SignalP *P* = 1.0). Six of the 52 residues are coded for by a first exon, with the remaining 46 residues coded for by a second exon (Table 4.1). *Acp95EF* has genes in close proximity on both genomic flanks. Based on tBLASTn analysis, both of these genes are present in *D. pseudoobscura* (Fig. 4.11). The proximal neighbor, *CG13609*, generates a highly significant tBLASTn hit to a portion of *D. pseudoobscura* chromosome 4 (E = 3e-42). On the opposite flank and on the reverse strand, *CG5677* is also highly conserved in the same relative position in *D.*

pseudoobscura (E = 3e-96). tBLASTn analysis of *Acp95EF*, however, did not produce even a weak hit to any portion of the *D. pseudoobscura* genome. Conservation of Muller elements within *Drosophila* suggests *D. melanogaster* chromosome 3R is homologous to *D. pseudoobscura* chromosome 2 (Lakovaara and Saura 1982; Steinemann, Pinsker and Sperlich 1984). Whether this apparent 3R-to-4 homology is real or an error in the *D. pseudoobscura* genome assembly is unclear. Regardless, the microsynteny of *Acp95EF* flanking genes clearly defines a candidate region for a *D. pseudoobscura* ortholog.

The region of microsynteny defined by *CG13609*/*CG5677*, which would contain *D. pseudoobscura Acp95EF*, is only 204 bp, compared to 1.2 kb in *D. melanogaster*. The genomic sequence from start to stop codon of *D. melanogaster Acp95EF* spans 221 bp. Given the requirements for 5'- and 3'-UTRs, it seems highly improbable that a *D. pseudoobscura Acp95EF* homolog is located within this 204 bp *D. pseudoobscura* genomic sequence. The small size of the candidate region coupled with encroaching 3'-UTRs of *CG13609*/*CG5677* made reverse Northern analysis superfluous. Computational analysis is enough to dismiss the hypothesis of a microsyntenic *D. pseudoobscura* ortholog. There is only one possible initiation codon in this region. Unlike *D. melanogaster Acp95EF* (SignalP, $P = 1.0$), an intronless *D. pseudoobscura* peptide sequence originating from this codon is not strongly predicted to have a signal peptide (SignalP, $P = 0.71$) and could not exceed 23 residues. Furthermore, an ortholog of comparable length would be impossible within this region, even assuming intron loss in *D. pseudoobscura*. Given the requirements for intron splicing sites and conservatively assuming a minimum intron size of 40 bp, the longest possible *D. pseudoobscura*

ortholog could still only consist of 30 residues, less than 58% of the size of the relatively small *D. melanogaster Acp95EF* protein. A signal sequence for this candidate is also not strongly predicted (SignalP, $P = 0.64$). Thus, our computational evidence leads us to conclude that a *D. pseudoobscura Acp95EF* ortholog is not present within this microsyntenic region and that *Acp95EF* is likely a *D. melanogaster* orphan.

Acp98AB

Acp98AB is in a gene-rich portion of chromosome 3R in *D. melanogaster*. It is located within the 757 bp intron of *CG12879*. The *Acp98AB* ORF does not contain any easily detected signature sequences for computational analysis. There is no evidence of a typical methionine initiation codon and predicted peptide lengths vary from 28-31 residues, depending on the assumed first codon. There are no conserved domains and no evidence for a signal peptide sequence (SignalP, $P = 0.0$; Table 4.1). There are no tBLASTn hits in *D. pseudoobscura* to suggest an ortholog to *Acp98AB*. The neighboring genes, however, reveal the homologous region in *D. pseudoobscura*. tBLASTn scores for the second exon of *CG12879* (E = 1e-162), as well as two distal neighbors, *CG12876* and *CG12878* (E = 0.0 and 1e-111, respectively) clearly indicate this homologous region as a portion of *D. pseudoobscura* chromosome 2 (Fig. 4.12). This homology is also reinforced by BLASTn analysis of 2 kb of non-coding DNA proximal to *CG12879* in *D. melanogaster*. A total of seven small nucleotide sequences, averaging 58 bp in length, are microsyntenous between the two species (E values from E = 5e-24 to E = 3e-4;

partially depicted by homologous region a, Fig. 4.12). One additional gene, *CG12880*, is immediately proximal to these matching nucleotide sequences. tBLASTn analysis shows that this gene is also in a microsyntenic position in *D. pseudoobscura* (E = 2e-62, not shown in Fig. 4.12). Just 5' of *CG12878* CDS, BLASTn analysis identified one additional microsyntenic nucleotide sequence, depicted as region c in Fig.4.12 (E = 2e-12, 51/55 identical).

Comparison of the relative positions of these genes shows an inversion event between *D. melanogaster* and *D. pseudoobscura*. Based on clear regions of orthology, this inversion covers at least the second exon of *CG12879* and the entire *CG12876* gene. The regions labeled as a and c in Fig. 4.12 are the closest conserved markers clearly outside of the inversion breakpoints. The unknown location of the first *CG12879* exon in *D. pseudoobscura* (no tBLASTn or BLASTn identity was detected) complicates efforts to determine whether or not *Acp98AB* might have been included in the inversion. In fact, our RACE data show *CG12879* to be an intronless gene in *D. pseudoobscura*. There are no intron gaps in the consensus 5' *D. pseudoobscura* RACE sequence and a single ORF possibility (moving upstream from the putative initiation codon, a stop codon comes into frame before an alternative initiation codon is reached). The protein alignment between species is very robust beyond the missing *D. pseudoobscura* first exon, with the first *D. pseudoobscura* residue matching the 61st residue in *D. melanogaster* and high levels of conservation continuing to the end of the protein for an overall 69.8% level of similarity. We should note that there is no empirical support from full-length cDNAs or ESTs for the annotated *D. melanogaster* first exon. In fact, an alternate initiation codon exists in

D. melanogaster that leads to a 398 residue, single exon protein that is the exact same size as its *D. pseudoobscura* counterpart. Thus, we proceeded to target candidate regions in *D. pseudoobscura* under the conservative assumption that the first exon of *D. melanogaster CG12879* may not be real.

If *Acp98AB* were included in the inversion, we would expect the *D. pseudoobscura* ortholog to be on the minus strand between *CG12879* and conserved region c in Fig. 4.12. Alternatively, if *Acp98AB* were outside of the inversion breakpoints, we would expect the *D. pseudoobscura* ortholog to be on the plus strand between conserved region a and *CG12876* in Fig. 4.12. These possibilities lead to candidate regions of 352 bp and 2 kb, respectively. BLASTn analysis of the 2 kb sequence to all *D. melanogaster* sequences reveals a highly significant match to *Jonah99C* (four separate matches averaging 116 bp, E scores from 2e-55 to 1e-9; depicted as region b of Fig. 4.12), a member of a gene family that includes multiple repetitive sequences (Carlson and Hogness 1985). Excising the sequence spanning *Jonah99C* BLASTn matches, two *D. pseudoobscura* candidate regions of 797bp and 407 bp exist between microsyntenic region a and *CG12876*. The 407 bp candidate region can be further condensed to approximately 360 bp, considering the requirements for a *CG12876* 5' UTR. Thus, through our analyses of *D. melanogaster*/*D. pseudoobscura* micro-synteny, we have narrowed the *D. pseudoobscura Acp98AB* candidate space to three sequences of *D. pseudoobscura* chromosome 2, covering approximately 1.5 kb and spanning less than 7 kb.

Due to the fragmented nature of the candidate regions and the uncertainty about transcription boundaries of the tightly arranged adjacent genes, reverse Northern and RACE analyses were impractical. The power of our computational analyses was compromised by the short *Acp98AB* gene sequence, the lack of a traditional methionine start codon, and the absence of signature sequences such as a conserved domain or predicted signal sequence. A total of 19 ORFs are possible within the three *D. pseudoobscura* candidate sequences (13, 3, and 3 for the three candidate sequences from left to right; Fig. 4.12). However, none show any resemblance to *D. melanogaster Acp98AB*. Thus, we propose that *Acp98AB* is a *D. melanogaster* orphan, though a highly diverged *D. pseudoobscura* ortholog would be very difficult to detect.

Discussion

Of the 13 *D. melanogaster Acps* investigated here, four are clearly present in the expected homologous region of *D. pseudoobscura* (*Acp26Aa*, *Acp26Ab*, *Acp32CD* and *Acp53Ea*) while two, *Acp62F and Acp70A*, are likely present in *D. pseudoobscura*, but in non-homologous locations, perhaps as a result of transposition or other rearrangements between species. For seven of the 13 *Acps* (*Acp29AB*, *Acp33A*, *Acp36DE*, *Acp63F*, *Acp76A*, *Acp95EF* and *Acp98AB*) we have neither computational nor molecular support for the presence of a *D. pseudoobscura* ortholog. We cannot definitely state that all seven of these putative orphan *Acps* are absent from *D. pseudoobscura*, as it is always formally possible that gene absence is conflated with extremely high divergence and transposition to non-homologous locations. However, given our success in identifying

two cases of diverged *Acps* that are resident in non-homologous locations in the two species, we think it is likely that many, if not all, of the seven *Acps* in questions are absent from *D. pseudoobscura*. The most convincing case of an annotated *D. melanogaster Acp* that is absent from *D. pseudoobscura* is *Acp36DE*, due to its large size and the lack of sequence within the homologous microsyntenic region. Likewise, *Acp76A* is almost certainly absent from *D. pseudoobscura*. *Acp29AB* and *lectin-29Ca* are probably also *D. melanogaster* orphans, as other genes coding for serpin domains carry signature sequences that are easily detectable. We are less certain about *Acp33A*, *Acp63F*, *Acp95EF* and *Acp98AB*. Uncertainty about *Acp33A* is magnified by the incomplete assembly in this region. However, it is unlikely that *Acp63F*, *Acp95EF* and *Acp98AB* are located in their respective microsyntenic regions. Given the short lengths of these genes (their largest exons are 156 bp, 141 bp, and 96 bp, respectively), it is difficult to detect transposition combined with rapid evolution. *Acp70A* provides an example of the approximate limitations of our methods. We were able to identify the non-syntenic *D. pseudoobscura Acp70A* ortholog, despite its short length and limited tBLASTn similarity (E = 0.002). If any of the aforementioned putative orphans exist in *D. pseudoobscura*, they are likely to be non-syntenic and more diverged between species than *Acp70A*.

Varying levels of protein conservation were observed for the six genes for which homologs were identified in the two species. Of particular interest are proteins that are clearly orthologous based on genomic location, gene organization and length, and gene expression, but for which divergence is so great that protein sequences provide no

support for orthology. A good example is *Acp26Aa*, which is not detectable through tBLASTn analysis but is clearly orthologous in the two species. In *D. melanogaster*, *Acp26Aa* transferred during mating is processed by the female and has effects on oviposition during the first 24 hours postmating (Herndon and Wolfner 1995, Heifetz et al. 2000). Whether *Acp26Aa* has similar functions in the two species in spite of the lack of sequence similarity is an interesting question, which could be answered through genetic analysis of *D. pseudoobscura*.

Of the *Acps* that have been subjected to evolutionary analysis in *melanogaster* subgroup species, *Acp26Aa* shows the strongest evidence for directional selection, including Ka/Ks > 1 (Tsaur and Wu 1997), significant McDonald-Kreitman tests (Aguadé 1998; Tsaur, Ting, and Wu 1998), and overdispersed amino acid substitution (Kern, Jones, and Begun 2004). We were interested in determining whether the *D. pseudoobscura Acp26Aa* ortholog showed similar patterns of molecular polymorphism and divergence as those observed in the *melanogaster* subgroup. We collected population genetic data for *Acp26Aa* from *D. pseudoobscura* (six alleles) and its sister species, *D. persimilis* (one allele), along with a single outgroup species allele from *D. miranda*. There is evidence of gene flow between *D. pseudoobscura* and *D. persimilis*, though onset of divergence dates back approximately 589,000 years (Hey and Nielsen 2004). Our *Acp26Aa* data groups the single *D. persimilis* allele amongst the six *D. pseudoobscura* alleles. Thus, we report polymorphism and diverge data both with the *D. persimilis* allele included and removed from the data set (Tables 4.3-4). We found that the relative rates of replacement to silent site evolution in the *D. pseudoobscura*/*D.*

persimilis vs. *D. miranda* comparison are comparable to the rates of evolution in the *melanogaster* subgroup (Table 4.3). Replacement polymorphism in *D. pseudoobscura*/*D. persimilis* is similar to both African and American populations of *D. melanogaster*, whereas silent sites are more than twice as variable in *D. pseudoobscura*/*D. persimilis* and African *D. melanogaster* than American *D. melanogaster* (Table 4.3). Our McDonald-Kreitman test of *D. pseudoobscura*/*D. persimilis* vs. *D. miranda* sequences showed convincing evidence for adaptive protein evolution ($G = 5.76$, $P = 0.016$; Table 4.4). African *D. melanogaster* populations likewise show significant evidence of adaptive protein evolution ($P = 0.002$) while American *D. melanogaster* populations show a non-significant trend towards excess replacement fixations ($P = 0.109$), probably as a consequence of lower levels of polymorphism in this population. Thus, our data suggests *Acp26Aa* is evolving at comparable rates in both the *D. melanogaster* and *D. pseudoobscura* lineages and that adaptive protein evolution occurs in both lineages. The finding that *Acp26Aa* protein evolves rapidly in two distantly related *Drosophila* lineages as a result of directional selection suggests that a history of directional selection at this gene will be widely shared among species from this genus. It remains to be seen what other *Acps* or other types of proteins tend to be under directional selection during most of their evolutionary history. These data should also serve to remind us that low levels of protein similarity between species are as easily explained by directional selection as by lack of functional constraint. Note also that the *Acp26Aa* Ka/Ks ratio in *D. pseudoobscura*/*D. miranda* is about one, consistent with low functional constraint. However, the population genetics data provide much additional power to make inferences

about evolutionary mechanism. Given the long history of adaptive evolution between *D. melanogaster* and *D. pseudoobscura Acp26Aa*, a comparative functional analysis would be most interesting and could potentially reveal whether the underlying mechanisms of natural selection are similar in the two lineages.

Previous population genetic data from *Acp29AB* and *Acp36DE* support the idea that both have been under directional selection in *D. melanogaster*/*D. simulans* (Aguadé 1999, Begun et al. 2000). Thus, the fact that our analysis suggests that both are absent from the *D. pseudoobscura* genome is particularly interesting. There are two possible explanations for the presence/absence data. Either both genes were present in the *D. melanogaster*/*D. pseudoobscura* ancestor and then lost in the *D. pseudoobscura* lineage, or both were gained in the *D. melanogaster* lineage. The approaches used here, when applied to other *Drosophila* species, are likely to provide a clear answer to this question. Still, from an evolutionary perspective, either scenario is of interest. If the genes originated in the *D. melanogaster* lineage and are also under selection in that lineage, one might speculate that this is a common feature of lineage-specific new genes, consistent with data from other such genes in *Drosophila* (reviewed in Long et al. 2003). Alternatively, if the genes were lost in the *D. pseudoobscura* lineage but were under selection in *D. melanogaster*/*D. simulans*, the interpretation would be that radically different selection regimes had been operating in these two lineages.

Of course, the evolutionary questions have a parallel in issues relating to the functional biology of these two genes and these two species. For example, the evidence for directional selection of *Acp29AB* and *Acp36DE* in *D. melanogaster*/*D. simulans*

certainly suggests they are functionally "important." Though the function of *Acp29AB* is unknown, flies that are mutant for *Acp36DE* in *D. melanogaster* have major defects. *Acp36DE* protein is required for proper sperm storage. Females mated to mutant males lacking *Acp36DE* store only 15% as many sperm as females mated to wild-type males (Neubaum and Wolfner 1999). This protein binds to sperm heads and also localizes to the opening of the sperm storage organs (Bertram, Neubaum, and Wolfner 1996). The loss of sperm from seminal receptacles occurs rapidly on the second day after mating, thus affecting female patterns of remating as continued female resistance to male mating attempts requires stored sperm (Neubaum and Wolfner 1999). It would be fair to say that the *Acp36DE* protein plays an important role in *D. melanogaster* fertility. Given these data and our presence/absence data, there are two possible interpretations. Either the function of *Acp36DE* is required in both lineages, yet is fulfilled by another protein in *D. pseudoobscura*, or the functional biology of male-female interactions are sufficiently diverged such that not all functions are represented in all *Drosophila* lineages. Genetic analysis should allow these alternatives to be distinguished.

The ancestral *Drosophila* karyotype is five acrocentric rods (Ashburner 1989). In the *D. pseudoobscura* lineage, a relatively recent X-autosome fusion has resulted in a large X chromosome that contains roughly 40% of the genome, rather than the typical 20% for most species, including *D. melanogaster* (Powell and DeSalle 1995). In *D. melanogaster*, *Acps* and other genes associated with male reproduction appear to be underrepresented on the X chromosome (Wolfner et al. 1997, Parisi et al. 2003, Ranz et al. 2003). Conservation of *Drosophila* Muller elements strongly predicts that some *Acps*

that were on the chromosome corresponding to *D. melanogaster* 3L became X-linked in the lineage leading to *D. pseudoobscura* as a result of fusion of Muller elements (corresponding to X and 3L of *D. melanogaster*). If selection disfavors X-linked *Acps*, genes corresponding to 3L *Acps* in *D. melanogaster* should have been under strong selection for loss or transposition to an autosome in *D. pseudoobscura*. In fact, our two examples of *Acp*-related rearrangements leading to non-homologous locations for orthologs (*Acp62F* and *Acp70A*) were 3L-located *D. melanogaster* genes that have avoided XR-linkage in *D. pseudoobscura* (but see Stevison, Counterman, and Noor 2004 for XR-linked *Acps*). Moreover, two other *Acps*, *Acp63F* and *Acp76A*, that should be on XR in *D. pseudoobscura*, appear to be entirely absent from the *D. pseudoobscura* genome. Thus, none of the four *Acps* that should be X-linked in *D. pseudoobscura* as a result of an X-autosome fusion actually are X-linked. This supports the idea that X vs. autosome location can have major roles in the evolution of genome content and organization (Betrán, Thornton, and Long 2002).

One hypothesis for this pattern is that natural selection disfavors X-linked locations for male-advantage genes that are deleterious to females (Parisi et al. 2003). Our data are consistent with this hypothesis. *Acps* have been implicated as the likely components of seminal fluid that confer a cost of mating to females (Chapman et al. 1995). Little is known about the specific phenotypes associated with *Acp63F* and *Acp76A*. However, *Acp62F* is a protease inhibitor that is known to be toxic upon ectopic expression in females (Lung et al. 2002). *Acp70A*, though not shown to be deleterious to females, is a protein that serves a male agenda by increasing egg laying rate and reducing

female receptivity to re-mating (Chen et al. 1988, Chapman et al. 2003, Liu and Kubli 2003). Further analysis of comparative genomic data and elucidation of additional *Acp* phenotypes will help explain the X vs. autosome disparity in male-biased genes.

Figures

Figure 1.1. Comparison of replicate quantitative PCR scores. Each point represents a pair of replicate ΔC_T scores. Perfect replication would generate slope and R^2 scores of 1.0.

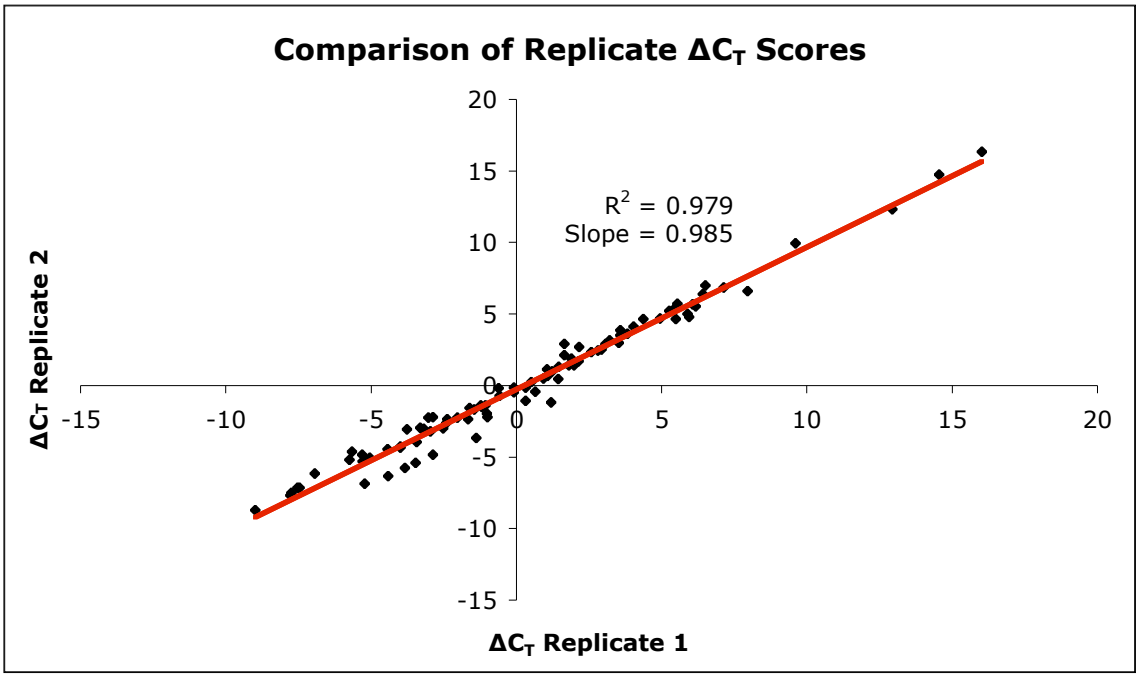


Figure 1.2. Alignment of *D. melanogaster* and *D. mojavensis* microsynteny around the *Tes100*/*115* gene region. Arrows next to genes indicate 5' to 3' orientation and dotted arrows between microsyntenic regions indicate significant BLAST similarity.

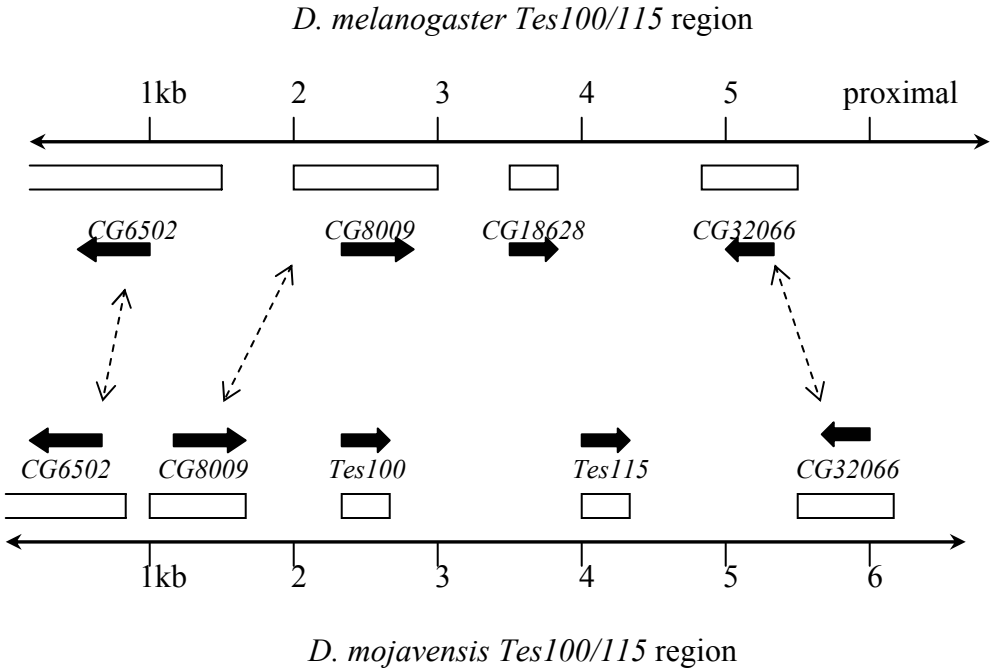


Figure 1.3. Correlation between absolute levels of expression and degree of tissue-specificity. The more tissue-specific genes (high $2^{-\Delta\Delta C_T}$) also tend to show higher absolute levels of expression (low ΔC_T). Testis-expressed genes are indicated by black diamonds, *Acps* by red triangles, and *moj-* genes by blue circles.

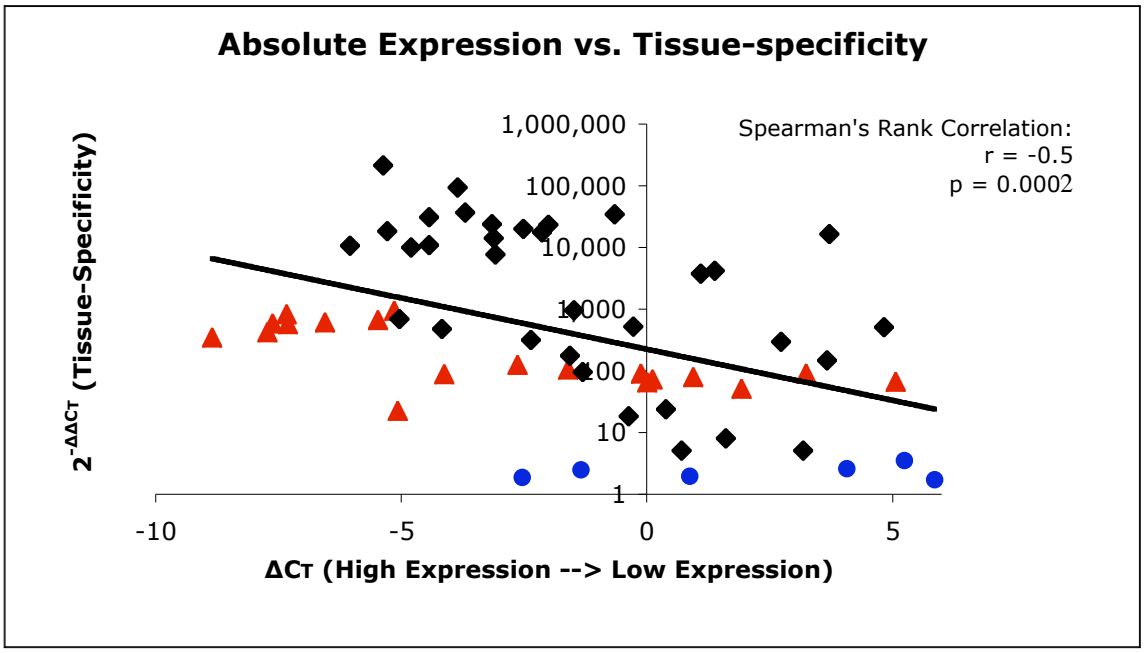


Figure 3.1. Phylogeny of *Acp5* duplicate genes. Evolution along each branch is shown as Ka/Ks values. The *D. mojavensis Acp5c* branch with values in bold has a Ka/Ks ratio that is significantly greater than one ($P < 0.05$).

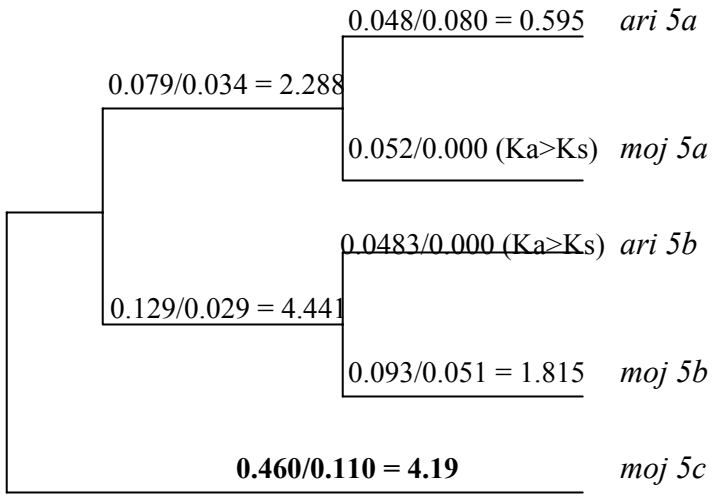


Figure 3.2. Phylogeny of *Acp16* duplicate genes. The branch with values in bold have a Ka/Ks ratio that is significantly greater than one ($P < 0.05$).

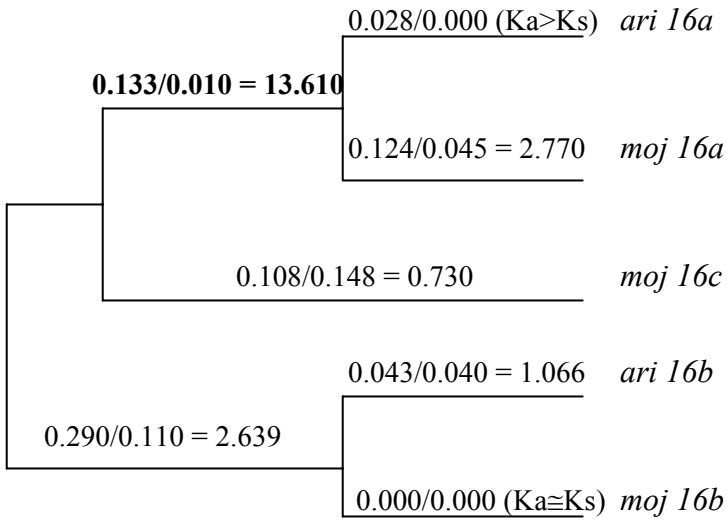


Figure 4.1. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp26Aa&Ab* gene region. *D. melanogaster Acps* are oriented with the 5' end to the left. The right side of *D. melanogaster* chromosomal regions are labeled "proximal' or "distal" to orient sequences with respect to centromeres. *D. pseudoobscura* chromosomal regions with rounded rather than arrowed ends depict contig endpoints from the incomplete genome assembly. Genes are represented by open rectangles, with no breaks for introns except for cases in which higher resolution is necessary. Solid horizontal arrows depict the 5' to 3' orientation of genes. Dashed arrows between *D. melanogaster*/*D. pseudoobscura* chromosomal segments depict homologous sequence as determined by BLAST analysis. Dotted horizontal lines indicate intergenic sequences that produce significant BLASTn results.

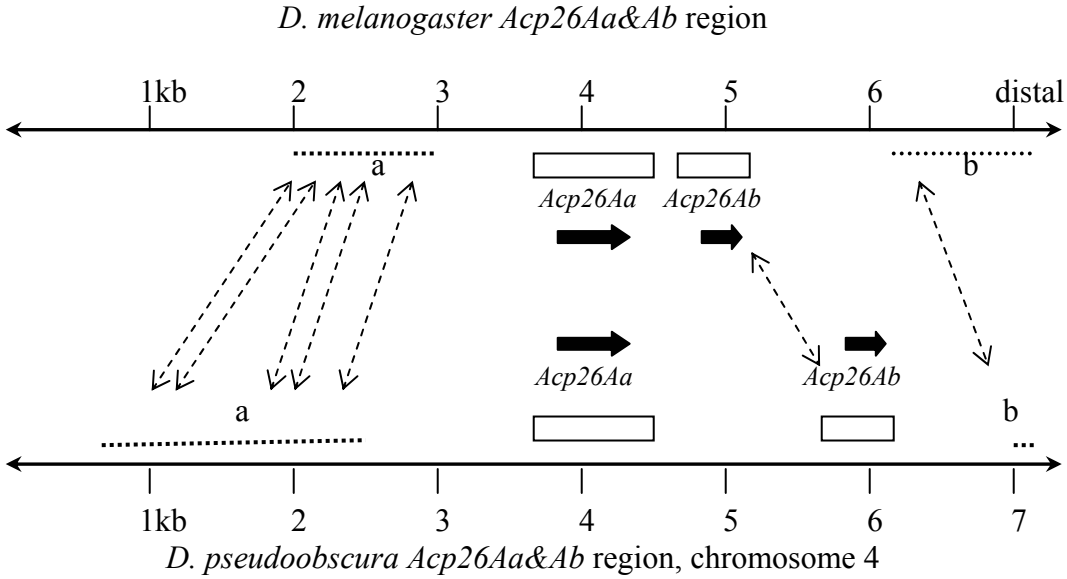


Figure 4.2. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp32CD* gene region.

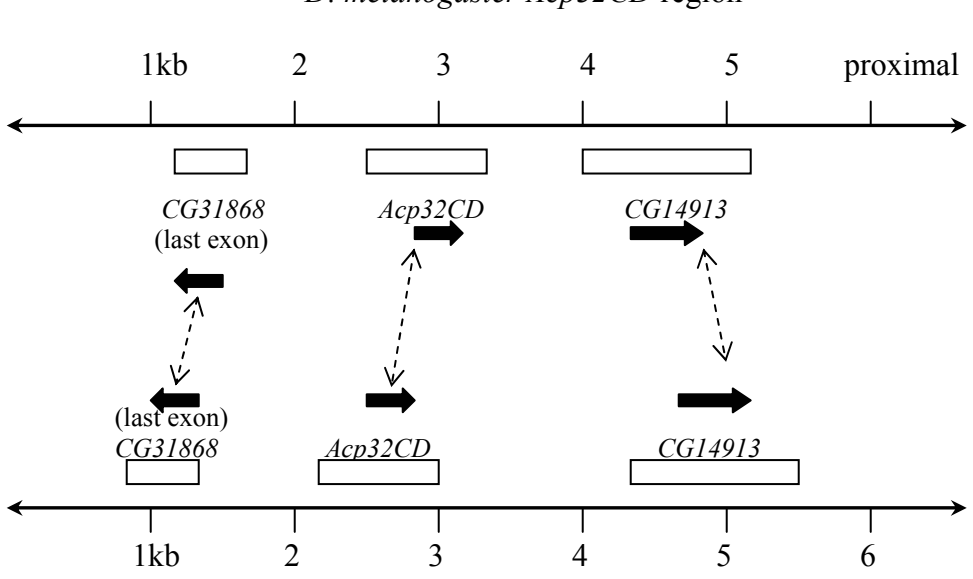


Figure 4.3. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp53Ea* gene region.

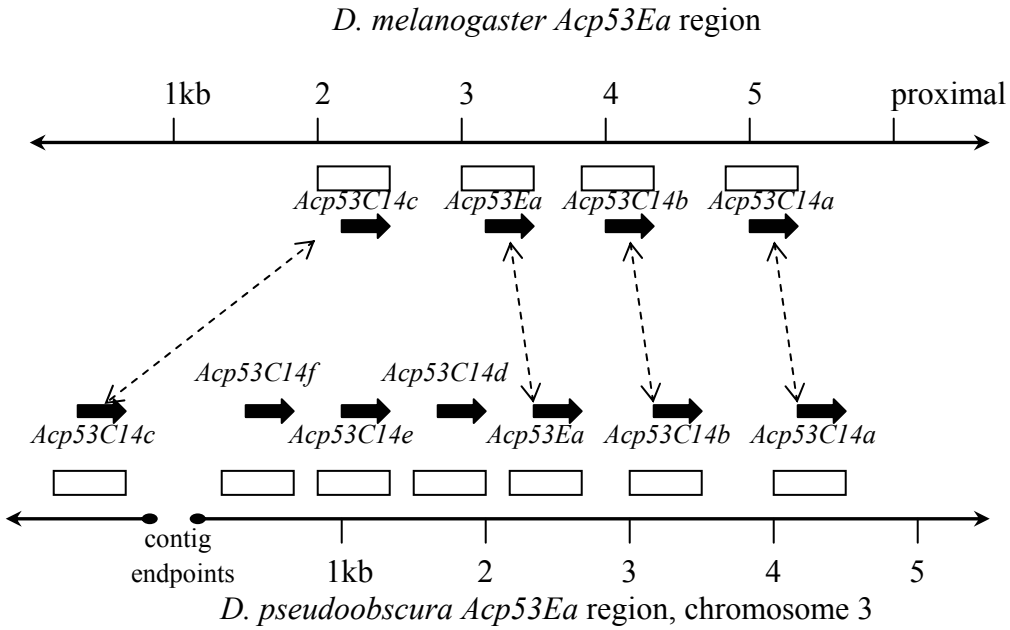


Figure 4.4. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp62F* gene region.

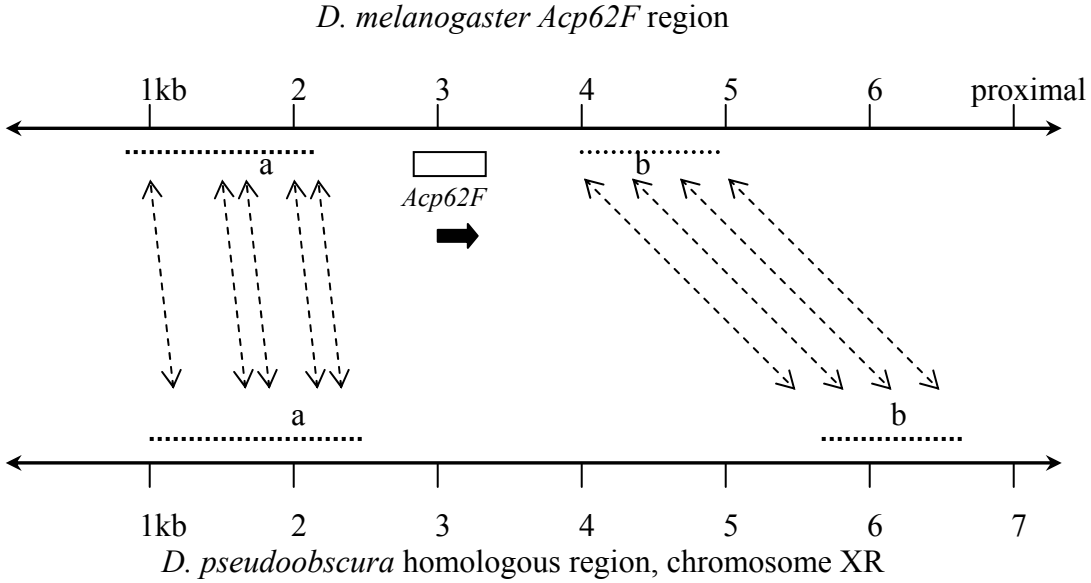


Figure 4.5. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp70A* gene region.

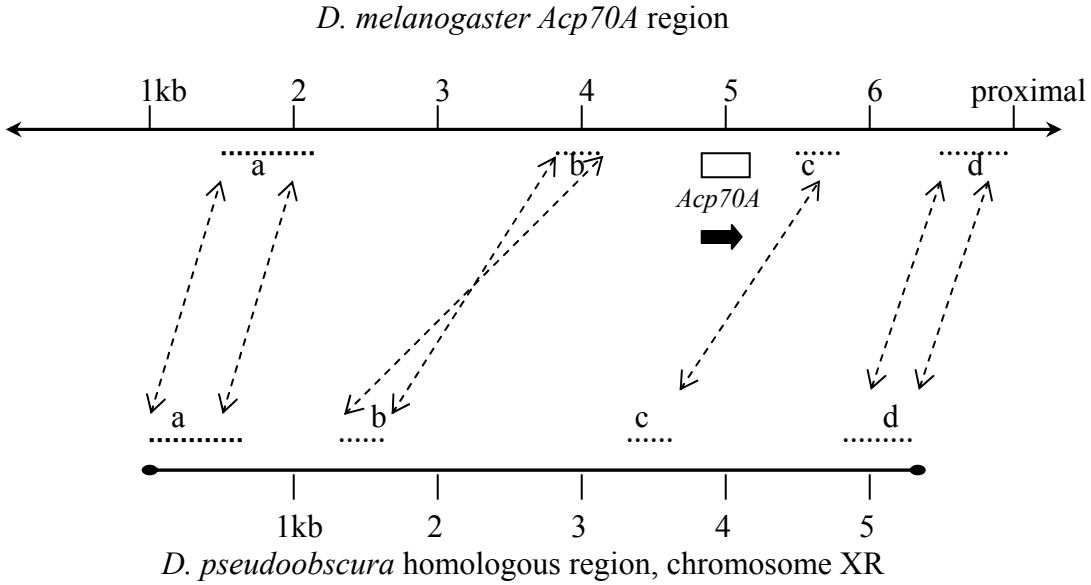


Figure 4.6. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp29AB* gene region.

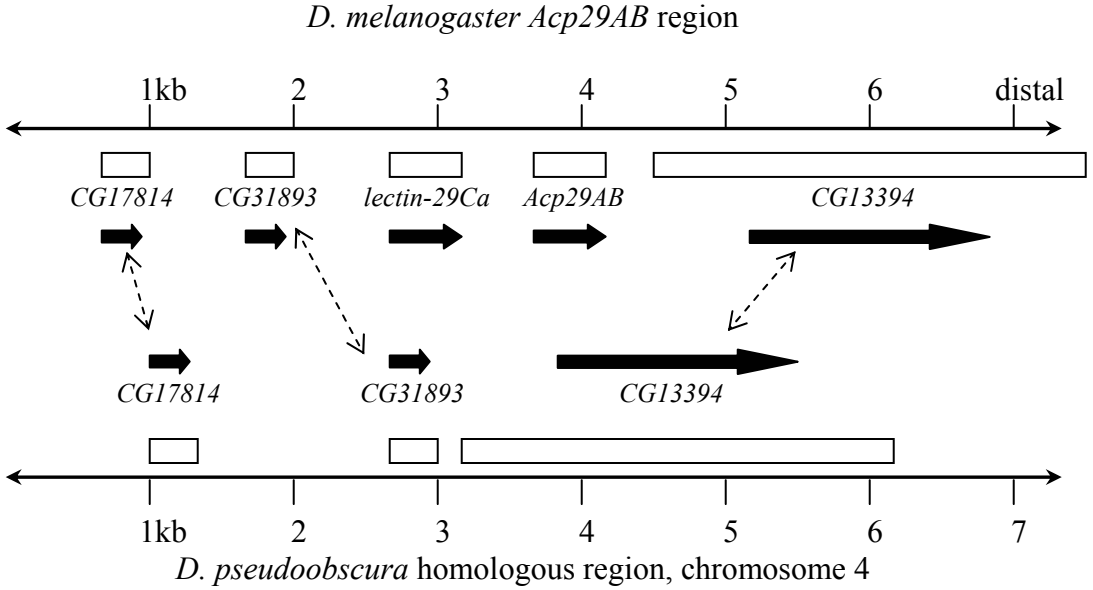


Figure 4.7. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp33A* gene region.

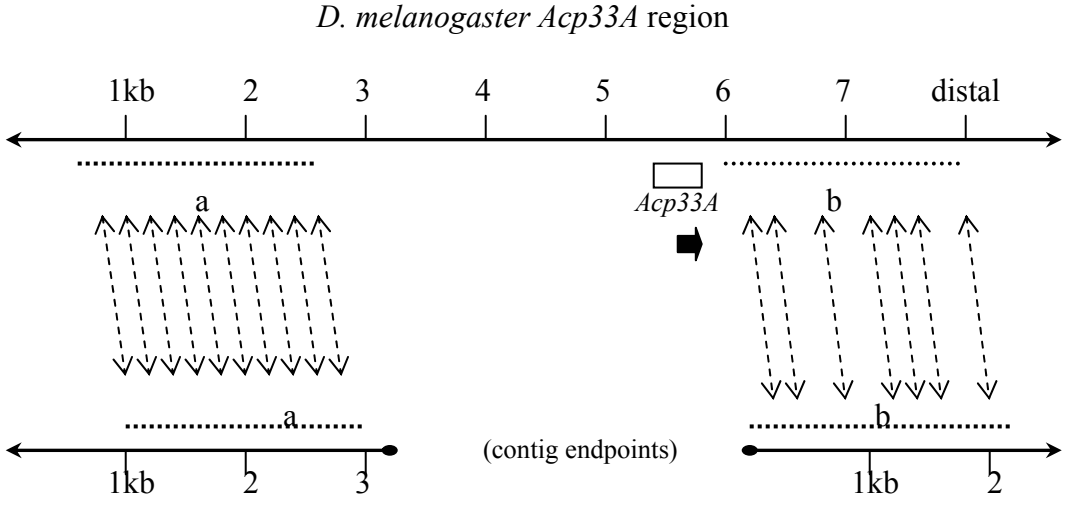


Figure 4.8. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp36DE* gene region.

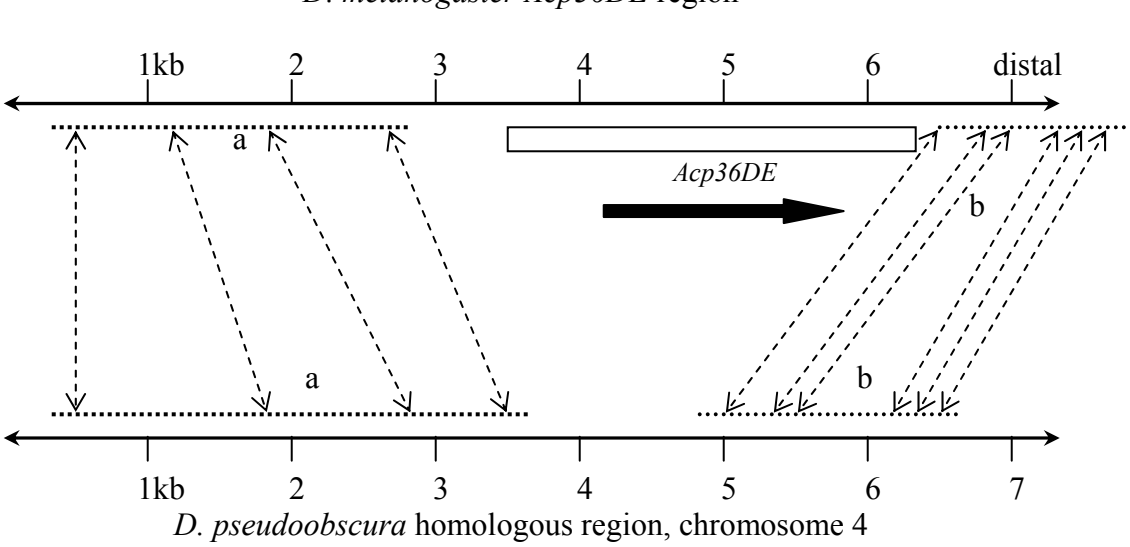


Figure 4.9. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp63F* gene region.

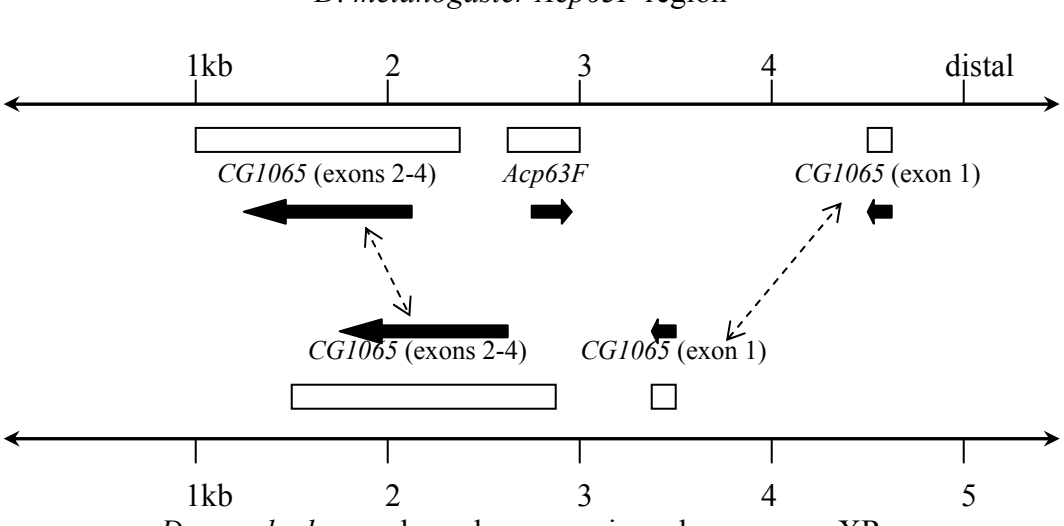


Figure 4.10. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp76A* gene region.

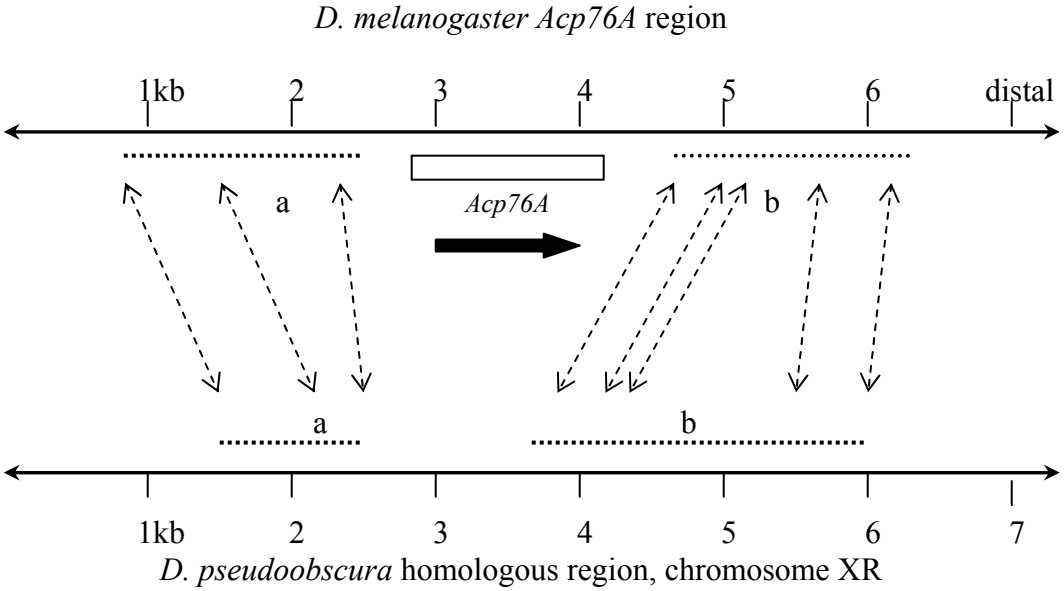


Figure 4.11. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp95EF* gene region.

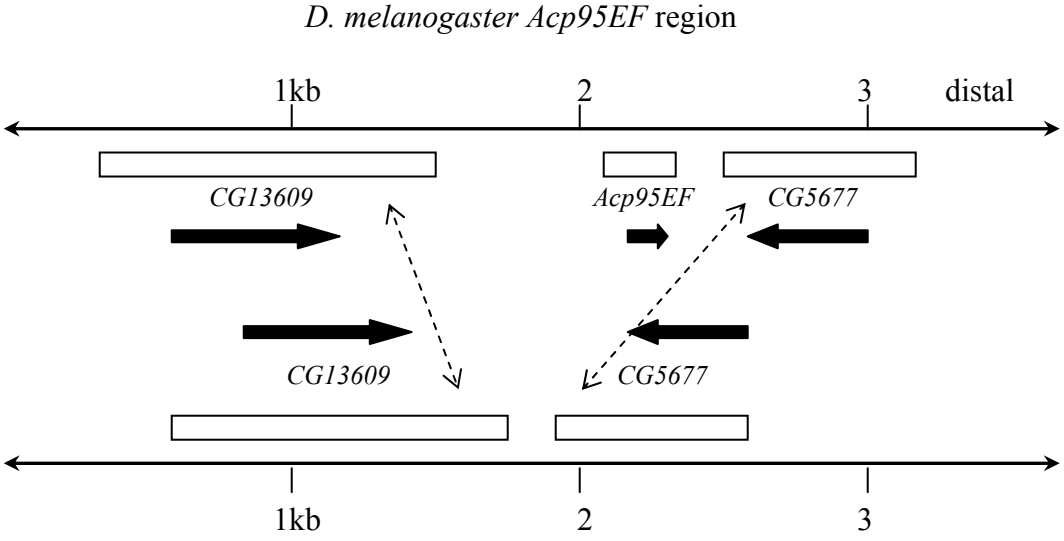


Figure 4.12. Alignment of *D. melanogaster* and *D. pseudoobscura* microsynteny around the *Acp98AB* gene region.

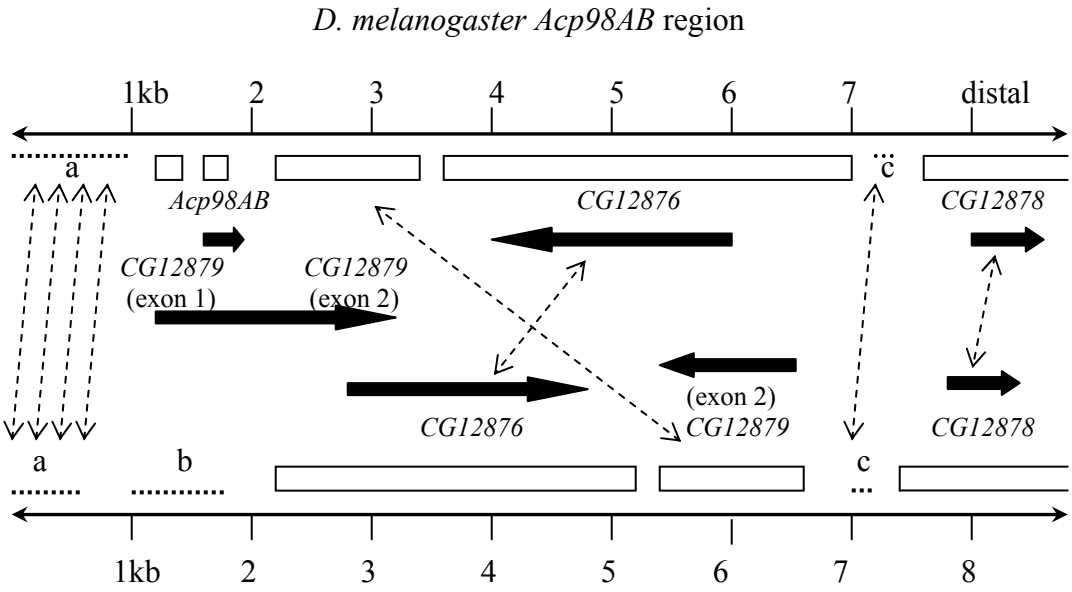
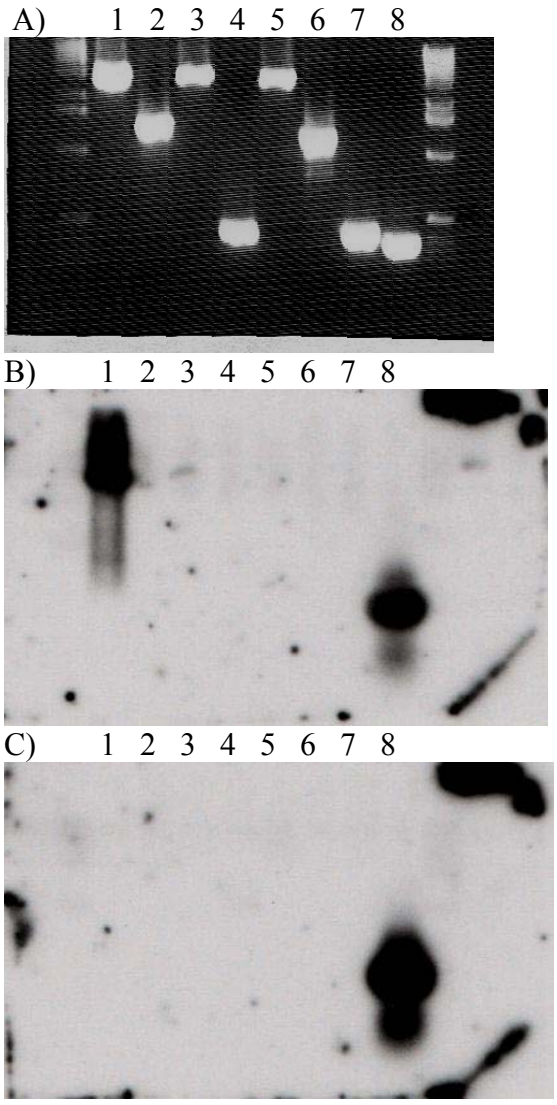


Figure 4.13. Reverse Northern of *D. pseudoobscura* ortholog candidate (i.e., microsyntenic) regions. PCR products spanning complete candidate microsyntenic regions and two control sequences were amplified from *D. pseudoobscura* genomic DNA, blotted, and probed with ^{32}P-labeled *D. pseudoobscura* cDNAs. A, photograph of ethidium gel exposed to UV light. B, blot probed with *D. pseudoobscura* male derived cDNA. C, blot probed with *D. pseudoobscura* female-derived cDNA. Lanes 1-6 correspond to microsyntenic regions for *Acp26Aa*, *Acp36DE*, *Acp62F*, *Acp63F*, *Acp70A*, and *Acp76A*, respectively. Lane 7 is the *pse-CG12880* intronic region negative control. Lane 8 is the *pse-CG7808* ribosomal protein positive control.



Tables

Table 1.1. EST Distribution of the *D. mojavensis* Male Reproductive Tract cDNA Library

Gene/EST	No. Clones	Before Multiplex	After Multiplex
Acp1	36	35	1
Acp2	5	4	1
Acp3	9	6	3
Acp4	9	8	1
Acp5a	36	27	9
moj6	2	2	
Acp7	6	5	1
Acp8	4	4	
moj9	1	1	
moj10	1	1	
Acp11	1	1	
moj12	1	1	
moj13	2	2	
Tes14	1	1	
Acp15	11	7	4
Acp16a	1	1	
Acp17	25	18	7
moj18	1	1	
Acp19	1	1	
moj20	2	1	1
Acp21a	7	5	2
Acp22	1	1	
Acp23	3	3	
Acp24	1	1	
Acp25	1	1	
Acp27b	1	1	
Acp27a	1		1
moj28	1		1
moj29	2		2
moj30	1		1
Tes31	1		1
moj32	1		1
Tes33	1		1
moj34	2		2
moj35	1		1

Table 1.1. Contd.

Gene/EST	No. Clones	Before Multiplex	After Multiplex
Acp36	1		1
moj37	1		1
moj38	1		1
Tes39	1		1
Tes40	1		1
Tes41	1		1
Acp42	1		1
moj43	1		1
moj44	1		1
Acp45	1		1
moj46	1		1
moj47	1		1
Acp48	1		1
moj49	1		1
moj50	2		2
moj51	1		1
moj52	1		1
moj53	1		1
Totals Clones:	199	139	60

Table 1.2. EST Distribution of the *D. mojavensis* Male Testis cDNA Library

Genes/ESTs	No. Clones X Unique ESTs
Tes100	20 X 1
Tes115	9 X 1
Tes110	4 X 1
Tes118, Tes154	3 X 2
Tes101, Tes104, Tes107, Tes108, Tes111, moj117, Tes134, moj157, moj163	2 X 9
moj102, moj103, Tes105, Tes106, Tes109, Tes112, Tes113, Tes114, moj116, Acp119, Tes120, moj121, Tes122, Tes123, Tes124, moj125, moj126, Tes127, moj128, Tes129, Tes130, Tes131, moj132, Tes133, moj135, moj136, moj137, moj138, moj139, Tes140, moj141, moj142, moj143, moj144, moj145, moj146, moj147, moj148, moj149, moj150, moj151, moj152, moj153, moj155, moj156, moj158, moj159, moj160, moj161, moj162, moj164, moj165, moj166, moj167, moj168, moj169, moj170, moj171, moj172, moj173, moj174, moj175, moj176, moj177, moj178, moj179, moj180, moj181, moj182, moj183, moj184, moj185, moj186, moj187, moj188, moj189, moj190, moj191, moj192, moj193, moj194, moj195, moj196, moj197, moj198, moj199, moj200, moj201, moj202, moj203, moj204, moj205, moj206, moj207, moj208, moj209, moj210, moj211, moj212, moj213, moj214, moj215, moj216, moj217, moj218	1 X 105
Total Clones Sequenced:	162

Table 1.3. BLAST and Conserved Domain Analysis of *D. mojavensis* ESTs

Gene/EST	Top BLAST Hit[a]	E Score	Secondary Hit[b]	Testis EST?[c]	Conserved Domain?[d]
Acp1	*CG15616-PA Acp53C14b*	7E-03	----	no	----
Acp2	*CG8626-PA Acp53C14a*	8E-08	2E-06	yes	----
Acp4	*CG11395-PA*	8E-08	----	no	----
Acp19	*CG9540-PA Ag5r*	5E-11	1E-10	yes	SCP
Acp25	*CG8622-PA Acp53Ea*	4E-03	----	no	----
Acp27a	*CG1385-PA Def*	5E-03	----	no	----
Acp36	**CG16713-PA*	2E-25	4E-19	yes	Kunitz family of serine protease inhibitors
Acp48	*CG12172-PA Spn43Aa*	2E-03	----	no	Serpin (serine protease inhibitor)
Tes14	*unannotated (protein)	1E-25	----	no	----
Tes31	**CG4523-PB*	2E-25	----	yes	----
Tes33	*CG17210-PA*	1E-103	1E-103	yes	SCP
Tes39	**CG3450-PA*	9E-39	----	no	Ubiquitin
Tes40	**CG9828-PA*	2E-56	3E-18	yes	DnaJ-class molecular chaperone with C-terminal Zn finger domain
Tes41	**CG5968-PA*	5E-32	4E-04	yes	----
Tes101	*CG14926-PA*	6E-03	----	yes	----
Tes104	*CG5106-PA*	1E-114	1E-114	yes	SCP
Tes105	**CG16972-PA*	6E-14	----	no	----
Tes106	**CG30334-PA*	7E-10	----	yes	----
Tes107	**CG31740-PA*	2E-06	----	yes	----
Tes109	*CG6209-PA*	7E-04	3E-03	yes	----
Tes110	**CG15219-PA*	3E-05	----	yes	----
Tes111	**CG31226-PB*	2E-07	5E-04	yes	----
Tes114	*CG5144-PA*	3E-10	1E-07	yes	ATP: guanido phosphotransferase
Tes118	*unannotated (protein)	3E-12	----	no	----
Tes120	*CG5024-PA*	3E-49	1E-41	yes	EF-Hand superfamily
Tes122	**CG7625-PA VhaM9.7-2*	3E-44	2E-20	yes	----
Tes123	*CG8174-PC SRPK*	6E-13	5E-12	yes	Serine/Threonine protein kinases
Tes124	**CG14079-PA*	1E-14	----	no	----
Tes127	**CG10090-PA Tim17a1*	2E-07	6E-05	no	----
Tes129	**CG3708-PA*	4E-32	3E-18	yes	Nucleosome assembly protein
Tes131	**CG4218-PA*	3E-25	----	yes	----

Table 1.3. Contd.

Gene/EST	Top BLAST Hit[a]	E Score	Secondary Hit[b]	Testis EST?[c]	Conserved Domain?[d]
Tes133	**CG14346-PA*	1E-43	----	no	----
Tes134	**CG33189-PA*	3E-21	3E-04	yes	----
Tes140	**CG12163-PD*	3E-16	----	yes	----
Tes154	**CG10252-PA*	9E-61	4E-06	yes	----
moj9	**CG5210-PA Chit*	9E-87	3E-57	no	Chitinase
moj10	*CG8495-PA*	1E-30	----	yes	Ribosomal protein S14
moj12	*CG7808-PD*	8E-79	----	yes	Ribosomal protein S8e
moj18	*CG4087-PA RpP2*	9E-33	----	yes	Ribosomal protein L12E/L44/L45/RPP1/RPP2
moj28	unannotated (nucleotide)	1E-139	----	yes	----
moj29	**CG2852-PA*	3E-84	3E-37	yes	pro isomerase
moj30	**CG8460-PA*	9E-72	----	no	Glycosyl hydrolase
moj32	**CG3654-PD*	8E-32	----	no	----
moj34	*CR40456-RA 18SRNA*	3E-47	----	yes	N/A
moj37	*CG6113-PA*	5E-82	6E-54	no	ab-hydro lipase
moj38	*CG8332-PA*	2E-63	----	no	Ribosomal protein S19
moj43	*CG3922-PB RpS17*	1E-53	----	no	Ribosomal protein S17E
moj44	*CG1652-PA lectin-46Cb*	1E-67	2E-43	yes	C-type lectin
moj46	*CG15168-PA*	2E-23	----	no	----
moj49	*CG9538-PA Ag5r*	3E-45	2E-39	yes	SCP
moj50	*CR40456-RA 18SRNA*	3E-83	----	yes	----
moj51	*CG6105-PA*	6E-48	2E-24	yes	Mitochondrial ATP synthase g subunit
moj52	*CG14708-PA*	2E-20	3E-13	no	----
moj53	*CR40456-RA 18SRNA*	1E-148	----	yes	N/A
moj102	unannotated (protein)	3E-10	----	yes	----
moj116	*CG6372-PA*	9E-53	7E-50	yes	Leucyl aminopeptidase
moj117	unannotated (nucleotide)	1E-131	----	no	N/A
moj125	*CG5048-PA*	2E-33	----	yes	----
moj132	*CG4651-PA RpL13*	3E-92	----	no	Ribosomal protein L13E
moj135	unannotated (nucleotide)	6E-06	----	no	N/A
moj137	**CG6773-PA sec13*	1E-41	3E-17	no	WD40 domain
moj138	*CG32267-PA*	8E-04	----	yes	----
moj139	*CG8138-PA*	5E-04	----	no	----
moj143	*CG31916-PA*	1E-25	2E-05	no	----
moj145	*CG1913-PA alphaTub84B*	2E-94	3E-94	yes	Tubulin/FtsZ
moj147	*CG2980-PA*	3E-56	----	no	----
moj148	*CG11840-PA shanti*	4E-26	----	yes	----

Table 1.3. Contd.

Gene/EST	Top BLAST Hit[a]	E Score	Secondary Hit[b]	Testis EST?[c]	Conserved Domain?[d]
moj150	*CG15369-PA*	6E-26	2E-21	yes	Cystatin-like
moj151	*CG12438-PA*	2E-20	----	no	----
moj152	**CG9941-PA*	1E-18	----	no	----
moj161	*CG1827-PA*	7E-21	2E-13	no	Asparaginase
moj162	*CG5614-PA*	2E-12	----	yes	LisH, Lissencephaly type-1-like homology motif
moj164	unannotated (nucleotide)	2E-12	----	yes	N/A
moj165	*CG17567-RB*	8E-04	7E-03	no	N/A
moj166	*CG8189-PA ATPsyn-b*	6E-75	1E-17	yes	Mitochondrial ATP synthase B chain precursor (ATP-synt B)
moj167	unannotated (protein)	5E-04	----	no	----
moj171	*CG8006-PA*	5E-23	2E-10	yes	----
moj173	*CG13245-PA*	8E-22	----	yes	----
moj174	*CG9007-RA*	2E-03	----	no	N/A
moj175	*CG4692-PA*	5E-59	1E-12	yes	----
moj177	*CG1728-PA Tim8*	2E-43	6E-08	yes	Tim10/DDP family zinc finger
moj180	*CG31345-PA*	2E-65	7E-57	yes	EF-hand, calcium binding motif
moj181	*CG8397-RA*	1E-05	----	no	N/A
moj182	*CG8309-PA*	4E-04	----	no	N/A
moj184	*CG13917-PA*	3E-36	----	no	----
moj186	*CG8989-PB His3.3B*	2E-71	2E-71	yes	Histone H3
moj188	*CG14724-PB CoVa*	1E-71	----	yes	Cytochrome c oxidase subunit Va
moj189	*CG4750-PA*	4E-46	5E-44	yes	Peptidase M17
moj191	*CG8226-PA*	7E-07	----	yes	----
moj193	*CG11314-PA*	4E-18	2E-17	no	ML (MD-2-related lipid-recognition)
moj194	*CG5273-PB*	2E-05	5E-04	no	N/A
moj195	*CG15693-PA RpS20*	9E-63	----	yes	Ribosomal protein S10p/S20e
moj196	*CG5762-PA*	9E-32	----	yes	----
moj199	*CG13364-RA*	2E-27	----	no	----

Table 1.3. Contd.

Gene/EST	Top BLAST Hit[a]	E Score	Secondary Hit[b]	Testis EST?[c]	Conserved Domain?[d]
moj201	*CG14684-PA*	2E-14	----	no	----
moj204	*CG14648-PA*	1E-05	----	no	----
moj205	*CG11858-PA*	1E-16	----	no	N/A
moj206	*CG17736-PA*	1E-07	----	yes	N/A
moj208	*CG2034-PA*	1E-22	----	yes	N/A
moj209	*CG5184-PA mRpS11*	4E-85	5E-06	no	RpsK, Ribosomal protein S11
moj213	*CG13601-PA*	3E-08	5E-03	yes	----

[a]Unannotatd BLAST matches are characterized as protein or nucleotide, as dictated by the type of sequence that returned the lowest E score. Putative orthologous matches are indicated by an * (see text for the subset of genes that were scrutinized for orthology).

[b]Indicates whether or not BLAST searches returned more than one homologous sequence.

[c]Indicates significant match (E < 1e-04) to *D. melanogaster* testis EST database (Andrews et al. 2000).

[d]The NCBI CDD database (Marchler-Bauer et al. 2003) was used for conserved domain analysis. Sequences corresponding to no known protein or highly truncated proteins are designated "N/A" as no CDD analysis was performed.

Table 1.4. Quantitative PCR Data for *D. mojavensis* and Related *D. melanogaster* Genes

Gene	ΔC_T	Fold Difference Relative to Second Most Abundant Tissue Template[a]			Tissue Order[b]
		$2^{-\Delta\Delta CT}$	Third Tissue	Least Abundant	
*Acp1**	-7.31	566.18	3.4E-03	2.2E-04	ATCF
*Acp2**	-5.14	932.81	3.1E-03	----	ATC
*Acp3**	-8.85	348.40	2.7E-03	9.6E-05	ATCF
*Acp5a**	-7.34	817.39	2.3E-03	1.7E-04	ATCF
*Acp7**	-7.73	416.40	1.3E-02	1.9E-03	ATCF
*Acp8**	-1.60	104.29	2.1E-03	----	ATC
*Acp11**	-7.62	578.07	4.4E-03	2.2E-04	ATCF
*Acp16a**	0.01	64.91	1.9E-03	----	ATC
*Acp19**	-6.55	602.63	4.6E-03	2.4E-03	ATFC
mel - CG9538[d,e]	1.89	6.36	1.2E-02	5.4E-03	CFTA
*Acp21a**	0.11	72.82	3.8E-03	----	ATC
*Acp22**	5.06	66.60	8.9E-02	2.7E-02	ATCF
*Acp24**	-0.12	90.51	2.5E-03	4.7E-04	ATCF
*Acp25**	1.92	50.81	----	----	AT
mel - Acp53Ea[d]	-1.24	47.54	1.2E-02	1.2E-04	ATCF
*Acp27a**	3.24	90.57	3.0E-02	----	ATC
mel - CG1385[f]	2.27	2.82	1.9E-02	5.3E-03	CFAT
*Acp27b**	0.95	78.10	8.0E-02	1.7E-02	ATCF
*Acp42**	-4.13	87.95	1.5E-03	6.1E-05	ATCF
*Acp45**	-5.47	663.23	8.1E-03	----	ATC
*Acp48**	-5.07	22.27	1.9E-04	----	ATC
mel - CG12172[d,e]	6.76	13.63	3.5E-01	2.4E-01	TFAC
*Acp119**	-2.64	124.86	----	----	AT
*moj9**	-2.54	1.88	1.9E-01	1.7E-01	TCFA
mel - CG5210[c,d,e]	3.71	2.15	4.2E-01	5.5E-02	CFTA
*moj29**	-1.34	2.52	6.4E-01	5.1E-01	ATCF
mel - CG2852[c,d,e]	3.56	1.94	9.1E-01	4.5E-01	TAFC
*moj30**	5.86	1.72	3.5E-01	2.5E-01	ATCF
mel - CG8460[c,e]	6.40	1.74	7.0E-01	3.2E-01	TAFC
moj32	4.07	2.61	6.1E-01	5.2E-01	TFAC
mel - CG3654[c]	14.64	2.53	4.4E-01	1.2E-01	FCTA
moj137	0.87	1.97	3.1E-01	3.0E-01	TACF
mel - CG6773[c,d,e]	9.77	1.21	5.3E-01	4.9E-01	ATCF

Table 1.4. Contd.

Gene	ΔC_T	Fold Difference Relative to Second Most Abundant Tissue Template[a] $2^{-\Delta\Delta CT}$	Third Tissue	Least Abundant	Tissue Order[b]
moj152	5.24	3.50	7.9E-01	6.8E-01	TFCA
mel - CG9941[c]	12.63	2.12	9.4E-01	1.3E-01	FTCA
Tes14	-0.38	18.40	7.2E-01	2.1E-01	TACF
mel - unannotated[c]	2.43	4.04	8.3E-01	5.5E-01	TCAF
mel - CG8446[g]	7.00	1.61	5.7E-01	1.2E-01	CTFA
Tes31	1.61	8.00	4.3E-01	2.4E-01	TCFA
mel - CG4523[c]	5.38	2.11	4.9E-01	3.9E-01	TCFA
*Tes33**	-5.29	18511.77	8.2E-01	3.5E-01	TCAF
mel - CG5106[d,e]	-1.59	2303.03	7.5E-01	4.5E-02	TCFA
Tes39	3.18	5.12	9.5E-01	7.9E-01	TCFA
mel - CG3450[c,e]	4.50	3.95	8.6E-01	8.4E-01	TAFC
Tes40	-1.30	96.73	5.4E-01	1.3E-01	TFCA
mel - CG9828[c,d,e]	5.63	14.89	2.8E-01	2.5E-01	TCAF
Tes41	1.10	3797.63	6.4E-01	2.6E-01	TCFA
mel - CG5968[c]	2.98	5252.83	1.0E+00	2.8E-01	TCAF
*Tes115**	-5.04	692.07	5.3E-03	9.0E-03	TACF
*Tes100**	-4.17	478.55	1.2E-02	6.5E-03	TACF
mel - CG18628[c]	-2.76	348.84	1.7E-01	5.3E-03	TAFC
Tes101	-3.69	36656.46	5.9E-01	----	TAC
mel - CG14926[f]	-2.55	2868.81	2.1E-01	9.1E-03	TACF
*Tes104**	-3.16	23873.76	6.3E-01	6.1E-01	TCAF
Tes105	-2.36	315.80	5.4E-01	3.0E-01	TFCA
mel - CG16972[c]	-0.40	74.59	4.0E-01	3.2E-01	TFCA
Tes106	-2.51	20104.52	3.8E-01	----	TAC
mel - CG30334[c]	1.89	207.88	4.6E-01	5.4E-02	TCFA
Tes107	-0.65	34588.46	7.4E-01	----	TAC
mel - CG31740[c]	1.11	43993.04	7.0E-01	----	TAC
Tes108	-3.85	92768.67	1.0E+00	2.1E-01	TACF
Tes109	-4.80	9915.91	9.9E-01	2.7E-01	TFAC
Tes110	-5.36	214980.12	1.0E+00	----	TAC
mel - CG15219[c]	-3.42	9052.73	3.7E-01	3.8E-02	TACF

Table 1.4. Contd.

Gene	ΔC_T	$2^{-\Delta\Delta CT}$	Fold Difference Relative to Second Most Abundant Tissue Template[a] Third Tissue	Least Abundant	Tissue Order[b]
Tes111	-4.43	10951.10	2.5E-01	2.3E-01	TACF
mel - CG31226[c]	0.09	14969.73	8.4E-02	----	TAC
Tes112	-4.44	30746.30	2.8E-01	1.4E-01	TACF
Tes113	-6.04	10789.56	6.2E-01	4.9E-01	TACF
Tes114	-1.49	954.76	8.9E-01	2.9E-01	TACF
Tes118	-2.01	23131.32	1.2E-01	----	TAC
Tes120	1.38	4182.25	----	----	TA
Tes122	0.71	5.10	6.2E-01	5.9E-01	TCAF
mel - CG7625[c,d]	7.28	2.88	8.5E-01	6.0E-01	TCAF
Tes123	-2.13	17695.50	3.7E-01	----	TCA
Tes124	4.82	514.99	1.0E+00	9.9E-01	TACF
mel - CG14079[c]	16.17	85.38	----	----	TF
Tes127	-1.56	177.69	6.8E-01	3.6E-01	TCAF
mel - CG10090[c,d]	5.45	9463.99	1.0E+00	6.2E-01	TACF
Tes129	0.39	23.80	3.6E-01	2.5E-01	TFCA
mel - CG3708[c,d,e]	5.38	3898.51	6.2E-01	4.9E-01	TACF
Tes130	2.73	298.98	9.9E-01	7.3E-01	TACF
Tes131	-3.11	14202.18	7.3E-01	3.0E-01	TCAF
mel - CG4218[c]	1.69	28882.82	4.0E-01	----	TAC
Tes133	3.71	16393.88	----	----	TF
mel - CG14346[c]	5.88	34022.18	1.0E+00	----	TAC
Tes134	3.67	147.97	4.7E-01	4.6E-01	TCFA
mel - CG33189[c]	2.45	1451.89	2.1E-01	1.1E-01	TCAF
*Tes140**	-0.28	525.25	9.8E-01	5.5E-01	TFAC
mel - CG12163[c]	2.62	772.72	8.1E-01	2.0E-01	TCAF
Tes154	-3.08	7690.17	2.2E-01	1.6E-01	TCAF
mel - CG10252[c,d]	-3.13	7667.65	2.4E-01	8.9E-03	TACF

*Indicates *D. mojavensis* genes with detected signal peptide sequences.

[a]The fold score of the second most abundant tissue is not shown since, by rule, it must always be equal to one.

Table 1.4. Contd.

[b]Tissues are listed according to levels of expression, most abundant tissue first. A = accessory gland, T = testis, C = male carcass (minus reproductive tracts), and F = whole female tissue.

[c]Putative *D. melanogaster* ortholog.

[d]Part of a Gene Family.

[e]Contains a shared protein domain.

[f]Too divergent to be certain about orthology.

[g]Gene corresponds to an alternate splice of the same genomic sequence, see text for details.

Table 2.1. Evidence of Genetic Differentiation Between *D. m. baja* and *D. m. mojavensis*

	D. m. baja vs. *D. m. mojavensis*			*D. m. baja*		*D. m. mojavensis*	
Gene	F_{ST}	Ka	Ks	$\pi_{rep.}$	$\pi_{syn.}$	$\pi_{rep.}$	$\pi_{syn.}$
Acp1	0.000	0.005	0.030	0.005	0.025	0.003	0.004
Acp2	0.038	0.019	0.018	0.019	0.031	0.015	0.012
Acp3	0.000	0.000	0.015	0.007	0.000	0.023	0.000
Acp5	-0.019	0.019	0.000	0.018	0.000	0.019	0.000
Acp7	0.864	0.020	0.006	0.000	0.000	0.002	0.011
Acp8	0.087	0.013	0.016	0.011	0.022	0.013	0.000
Acp16a	0.276	0.064	0.000	0.083	0.005	0.008	0.000
Acp16b	0.000	0.006	0.031	0.007	0.000	0.007	0.065
Acp19	-0.200	0.003	0.008	0.003	0.012	0.005	0.006
Acp21	0.407	0.032	0.007	0.026	0.000	0.010	0.014
Acp24	0.000	0.019	0.029	0.016	0.045	0.022	0.000
Acp25	0.190	0.003	0.012	0.000	0.014	0.003	0.009
Acp27	0.364	0.009	0.011	0.008	0.017	0.000	0.000
Acp42	0.370	0.003	0.039	0.002	0.025	0.005	0.008
Acp48	0.057	0.004	0.012	0.005	0.019	0.002	0.006
moj9	0.138	0.001	0.018	0.003	0.024	0.000	0.006
moj30	0.076	0.005	0.042	0.007	0.045	0.003	0.047
Tes14	0.118	0.000	0.012	0.000	0.000	0.000	0.024
Tes33	0.173	0.002	0.038	0.001	0.029	0.002	0.041
Tes100	-0.075	0.006	0.025	0.005	0.030	0.010	0.020
Tes101	-0.200	0.002	0.000	0.004	0.000	0.000	0.000
Tes104	0.370	0.000	0.016	0.000	0.014	0.000	0.009
Tes105	0.111	0.004	0.009	0.003	0.009	0.008	0.012
Tes106	0.133	0.005	0.048	0.006	0.027	0.004	0.031
Tes107	-0.060	0.000	0.021	0.000	0.027	0.000	0.021
Tes109	0.000	0.006	0.007	0.009	0.013	0.003	0.000
Tes110	0.338	0.003	0.000	0.003	0.000	0.000	0.000

Table 2.1. Contd.

	D. m. baja vs. *D. m. mojavensis*			*D. m. baja*		*D. m. mojavensis*	
Gene	F_{ST}	Ka	Ks	$\pi_{rep.}$	$\pi_{syn.}$	$\pi_{rep.}$	$\pi_{syn.}$
Tes112	0.056	0.000	0.032	0.000	0.032	0.000	0.032
Tes113	-0.182	0.002	0.023	0.002	0.030	0.003	0.021
Tes114	0.387	0.000	0.032	0.000	0.000	0.000	0.032
Tes115	0.000	0.002	0.000	0.000	0.000	0.004	0.000
Tes118	-0.030	0.002	0.010	0.001	0.018	0.005	0.000
Tes134	0.377	0.005	0.004	0.004	0.005	0.002	0.000
Tes154	0.314	0.003	0.019	0.001	0.016	0.000	0.022
all *Acps*	0.168	0.011	0.015	0.009	0.014	0.006	0.009
all *Tes*	0.144	0.003	0.013	0.002	0.015	0.002	0.014
all genes	0.150	0.006	0.015	0.005	0.016	0.004	0.013

Table 2.2. Polymorphism and Divergence at Individual *Acp*, *Tes*-, and *moj*- Genes

Gene	# alleles *a,mo,mu*[a]	# sites analyzed	ORF size	# coding analyzed	Sample	$\theta_{syn.}$	$\theta_{rep.}$	Ks	Ka	Ka/Ks
Acp1	7, 7, 1	326	354	288	*ari*	0.0000	0.0131	0.0463	0.0636	1.3744
					moj	0.0291	0.0056			
Acp2	7, 7, 1	237	354	234	*ari*	0.0218	0.0000	0.0638	0.0619	0.9705
					moj	0.0218	0.0184			
Acp3	7, 5, 0	305	207	150	*ari*	0.0342	0.0036	0.0799	0.0744	0.9316
					moj	0.0000	0.0168			
Acp5a	7, 7, 0	571	105	99	*ari*	0.0151	0.0057	0.1110	0.1099	0.9896
					moj	0.0000	0.0170			
Acp7	7, 7, 1	561	465	453	*ari*	0.0205	0.0086	0.0468	0.0378	0.8079
					moj	0.0068	0.0086			
Acp8	7, 7, 0	275	144	123	*ari*	0.0128	0.0179	0.1621	0.1214	0.7492
					moj	0.0128	0.0179			
Acp11	1, 1, 0	156	201	156		----	----	0.1600	0.0392	0.2450
Acp16a	7, 6, 0	151	189	141	*ari*	0.0000	0.0159	0.0596	0.1315	2.2049
					moj	0.0000	0.0299			
Acp16b	7, 4, 0	214	216	204	*ari*	0.0251	0.0184	0.0618	0.0499	0.8080
					moj	0.0336	0.0070			
Acp19	7, 7, 1	570	687+	510	*ari*	0.0107	0.0041	0.0267	0.0332	1.2424
					moj	0.0107	0.0031			
Acp21a	6, 7, 0	228	207	180	*ari*	0.0092	0.0066	0.0552	0.2274	4.1209
					moj	0.0086	0.0278			
Acp22	1, 2, 0	78	81	78		----	----	0.0000	0.0000	----
Acp24	6, 7, 0	135	129	120	*ari*	0.0000	0.0094	0.0559	0.0325	0.5825
					moj	0.0308	0.0175			
Acp25	7, 7, 1	324	354	294	*ari*	0.0346	0.0018	0.0582	0.0314	0.5386
					moj	0.0173	0.0018			
Acp27a	7, 7, 0	348	291	282	*ari*	0.0000	0.0019	0.0063	0.0135	2.1379
					moj	0.0120	0.0076			
Acp42	7, 7, 0	477	597+	363	*ari*	0.0104	0.0043	0.0724	0.0445	0.6146
					moj	0.0260	0.0043			
Acp45	1, 1, 0	372	408	372		----	----	0.0353	0.0323	0.9150
Acp48	7, 7, 0	516	630+	513	*ari*	0.0075	0.0040	0.1504	0.0861	0.5726
					moj	0.0187	0.0051			
Acp119	1, 1, 0	102	111	102		----	----	0.0000	0.0970	Ka>Ks
moj9	7, 7, 1	517	786+	447	*ari*	0.0228	0.0048	0.0495	0.0046	0.0938
					moj	0.0228	0.0024			
moj29	1, 1, 0	492	615	492		----	----	0.0374	0.0026	0.0695

Table 2.2. Contd.

Gene	# alleles *a*,*mo*,*mu*[a]	# sites analyzed	ORF size	# coding analyzed	Sample	$\theta_{syn.}$	$\theta_{rep.}$	Ks	Ka	Ka/Ks
moj30	7, 7, 1	631	621+	498	*ari*	0.0350	0.0043	0.0842	0.0056	0.0670
					moj	0.0455	0.0064			
moj32	1, 1, 0	180	429+	180		----	----	0.0000	0.0000	----
moj137	1, 1, 0	198	246+	198		----	----	0.0000	0.0000	----
moj152	1, 1, 0	303	396+	303		----	----	0.0893	0.0219	0.2452
Tes14	7, 7, 1	491	240	240	*ari*	0.0071	0.0000	0.0134	0.0000	0.0000
					moj	0.0153	0.0000			
Tes31	1, 1, 0	204	228	204		----	----	0.1280	0.0199	0.1555
Tes33	7, 7, 1	524	639+	468	*ari*	0.0606	0.0056	0.1169	0.0047	0.0401
					moj	0.0404	0.0022			
Tes39	1, 1, 0	210	219	210		----	----	0.0682	0.0000	0.0000
Tes40	1, 1, 0	393	505+	393		----	----	0.1217	0.0033	0.0271
Tes41	1, 1, 0	384	510	384		----	----	0.1274	0.0101	0.0793
Tes100	7, 7, 1	507	168	168	*ari*	0.0000	0.0153	0.0423	0.0273	0.6453
					moj	0.0353	0.0061			
Tes101	7, 7, 1	293	387	153	*ari*	0.0114	0.0000	0.0327	0.0012	0.0373
					moj	0.0000	0.0035			
Tes104	7, 7, 1	726	738+	663	*ari*	0.0239	0.0016	0.0725	0.0006	0.0077
					moj	0.0159	0.0000			
Tes105	7, 7, 1	363	234	231	*ari*	0.0145	0.0047	0.0206	0.0066	0.3185
					moj	0.0145	0.0047			
Tes106	7, 7, 1	368	207	207	*ari*	0.0184	0.0050	0.1611	0.0062	0.0383
					moj	0.0368	0.0050			
Tes107	7, 7, 1	501	126	126	*ari*	0.0389	0.0000	0.0815	0.0000	0.0000
					moj	0.0260	0.0000			
Tes109	7, 6, 0	234	927+	228	*ari*	0.0290	0.0132	0.0346	0.0311	0.8992
					moj	0.0000	0.0094			
Tes110	7, 7, 1	826	399	390	*ari*	0.0085	0.0014	0.0765	0.0029	0.0382
					moj	0.0000	0.0028			
Tes112	5, 7, 0	428	276	273	*ari*	0.0153	0.0000	0.0417	0.0048	0.1145
					moj	0.0325	0.0000			
Tes113	7, 7, 0	335	624	282	*ari*	0.0065	0.0037	0.0512	0.0072	0.1412
					moj	0.0194	0.0019			
Tes114	2, 7, 1	250	132+	96	*ari*	0.0000	0.0000	0.0633	0.0000	0.0000
					moj	0.0193	0.0000			
Tes115	6, 7, 1	321	204	207	*ari*	0.0000	0.0054	0.0448	0.0166	0.3706
					moj	0.0000	0.0025			

Table 2.2. Contd.

Gene	# alleles *a*,*mo*,*mu*[a]	# sites analyzed	ORF size	# coding analyzed	Sample	$\theta_{syn.}$	$\theta_{rep.}$	Ks	Ka	Ka/Ks
Tes118	4, 6, 0	729	936+	555	*ari*	0.0089	0.0076	0.0367	0.0151	0.4114
					moj	0.0142	0.0020			
Tes120	1, 1, 0	363	423+	363		----	----	0.0958	0.0106	0.1106
Tes122	1, 1, 0	267	267+	267		----	----	0.0172	0.0146	0.8488
Tes123	1, 1, 0	486	621+	486		----	----	0.1574	0.0768	0.4879
Tes124	1, 1, 0	159	651+	159		----	----	0.0277	0.0000	0.0000
Tes127	1, 1, 0	285	309+	285		----	----	0.0452	0.0282	0.6239
Tes129	1, 1, 0	405	525	405		----	----	0.0109	0.0032	0.2936
Tes130	1, 1, 0	150	174	150		----	----	0.0905	0.0125	0.1381
Tes131	1, 1, 0	528	603+	528		----	----	0.0407	0.0176	0.4324
Tes133	1, 1, 0	333	414+	333		----	----	0.0650	0.0160	0.2462
Tes134	7, 7, 1	805	609	558	*ari*	0.0238	0.0010	0.0540	0.0103	0.1897
					moj	0.0030	0.0039			
Tes140	1, 1, 0	240	240	240		----	----	0.0881	0.0169	0.1918
Tes154	7, 7, 1	696	579+	507	*ari*	0.0033	0.0011	0.0439	0.0019	0.0426
					moj	0.0263	0.0021			

[a]Number of alleles corresponding to *D. arizonae*, *D. mojavensis*, and *D. mulleri*, respectively.

Table 2.3. Polymorphism and Divergence of Gene Classes

		Polymorphism			Divergence[a]		
Gene Class	Sample	$\theta_{syn.}$	$\theta_{rep.}$	$\theta_{rep.}/\theta_{syn.}$	Ks	Ka	Ka/Ks
Acps	*ari*	0.0135	0.0066	0.4866	0.0643	0.0595	0.9257
	moj	0.0156	0.0093	0.5991			
Tes-	*ari*	0.0175	0.0037	0.2095	0.0682	0.0128	0.1873
	moj	0.0170	0.0025	0.1476			
moj-	*ari*	0.0292	0.0045	0.1553	0.0518	0.0060	0.1164
	moj	0.0346	0.0045	0.1308			
All genes	*ari*	0.0170	0.0049	0.2851	0.0650	0.0250	0.3842
	moj	0.0181	0.0053	0.2935			
sim Acps[b]		0.0280	0.0074	0.2643	0.1170	0.0497	0.4248
sim 3R[b]		0.0350	0.0013	0.0371	0.1080	0.0107	0.0991

[a]*D. simulans* genes divergence estimates are with respect to *D. melanogaster*.
[b]Data are from Begun et al. 2000

Table 2.4. Polarized *D. arizonae* vs. *D. mojavensis* divergence

Gene/Group	*D. arizonae*			*D. mojavensis*			Outgroup			
	Ka	Ks	Ka/Ks	Ka	Ks	Ka/Ks	Ka	Ks	Ka/Ks	Outgroup?
Acp1	0.0226	0.0139	1.6269	0.0480	0.0406	1.1808	0.1429	0.1616	0.8839	*D. mulleri*
Acp2	0.0366	0.0221	1.6559	0.0247	0.0300	0.8232	0.1513	0.2932	0.5160	*D. mulleri*
Acp5a	0.0714	0.0962	0.7426	0.0391	0.0000	Ka>Ks	0.2688	0.0400	*6.7231	*5b* duplicate
Acp7	0.0159	0.0245	0.6483	0.0275	0.0000	*Ka>Ks	0.2560	0.1200	2.1337	*D. mulleri*
Acp16a	0.0095	0.0244	0.3868	0.1538	0.0169	*9.1017	0.2708	0.1179	2.2972	*16c* duplicate
Acp16b	0.0406	0.0396	1.0248	0.0000	0.0000	----	0.5366	0.1905	*2.8161	*16a* duplicate
Acp19	0.0184	0.0167	1.0981	0.0163	0.0000	Ka>Ks	0.0953	0.0842	1.1313	*D. mulleri*
Acp25	0.0125	0.0458	0.2732	0.0207	0.0250	0.8265	0.1627	0.4233	0.3842	*D. mulleri*
Acp27a	0.0144	0.0000	Ka>Ks	0.0000	0.0134	0.0001	0.1100	0.0001	*Ka>Ks	*27b* duplicate
moj9	0.0029	0.0440	0.0653	0.0000	0.0298	0.0001	0.0145	0.0955	0.1516	*D. mulleri*
moj30	0.0000	0.0336	0.0001	0.0027	0.0498	0.0540	0.0109	0.1928	0.0564	*D. mulleri*
Tes14	0.0000	0.0152	0.0001	0.0000	0.0000	----	0.0186	0.1485	0.1254	*D. mulleri*
Tes33	0.0028	0.1064	0.0259	0.0028	0.0492	0.0574	0.0084	0.2142	0.0391	*D. mulleri*
Tes100	0.0000	0.0430	0.0001	0.0141	0.0420	0.3365	0.0219	0.2624	0.0836	*D. mulleri*
Tes101	0.0000	0.0000	----	0.0000	0.0191	0.0001	0.0102	0.0859	0.1191	*D. mulleri*
Tes104	0.0000	0.0302	0.0001	0.0000	0.0327	0.0001	0.0125	0.1529	0.0817	*D. mulleri*
Tes105	0.0000	0.0000	0.0000	0.0060	0.0000	Ka>Ks	0.0305	0.2418	0.1259	*D. mulleri*

Table 2.4. Contd.

	D. arizonae			*D. mojavensis*			Outgroup			
Gene/Group	Ka	Ks	Ka/Ks	Ka	Ks	Ka/Ks	Ka	Ks	Ka/Ks	Outgroup?
Tes106	0.0122	0.1532	0.0796	0.0000	0.0192	0.0001	0.0060	0.3648	0.0165	*D. mulleri*
Tes107	0.0000	0.0181	0.0001	0.0000	0.0179	0.0001	0.0000	0.0832	0.0001	*D. mulleri*
Tes110	0.0000	0.0000	----	0.0035	0.0630	0.0548	0.0139	0.0640	0.2173	*D. mulleri*
Tes114	0.0000	0.0611	0.0001	0.0000	0.0000	----	0.0264	0.0000	Ka>Ks	*D. mulleri*
Tes115	0.0162	0.0164	0.9889	0.0058	0.0166	0.3508	0.0702	0.0880	0.7979	*D. mulleri*
Tes134	0.0023	0.0356	0.0649	0.0098	0.0354	0.2760	0.0474	0.1407	0.3367	*D. mulleri*
Tes154	0.0000	0.0251	0.0001	0.0000	0.0233	0.0001	0.0082	0.1278	0.0640	*D. mulleri*
All *Acps*										
w/ *mul* only	0.0195	0.0235	0.8273	0.0257	0.0150	1.7163	0.1525	0.1798	0.8484	*D. mulleri*
w/ *mul* and dupl.	0.0220	0.0253	0.8715	0.0273	0.0131	2.0776	0.1801	0.1498	1.2024	*mul* + dupl.
All *Tes*	0.0020	0.0345	0.0578	0.0034	0.0306	0.1096	0.0199	0.1501	0.1326	*D. mulleri*
All *moj*	0.0014	0.0348	0.0407	0.0014	0.0382	0.0375	0.0130	0.1364	0.0951	*D. mulleri*

Note.—Ka/Ks ratios significantly greater than one ($P < 0.05$) are indicated by an asterisk.

Table 2.5. Individual Gene McDonald-Kreitman Tests

Gene	Polymorphic		Fixed		P^a
	Syn	Repl	Syn	Repl	
Acp1	5	10	1	10	0.130
arizonae	0	7	1	3	0.364
mojavensis	5	3	0	6	0.031*
Acp2	6	8	1	8	0.090
arizonae	3	0	0	5	0.018*
mojavensis	3	8	0	2	1.000
Acp3	3	5	2	6	0.589
arizonae	3	1	----	----	----
mojavensis	0	4	----	----	----
Acp5a	1	4	3	6	0.590
arizonae	1	1	3	3	1.000
mojavensis	0	3	0	2	----
Acp7	8	14	4	7	1.000
arizonae	6	7	2	3	0.813
mojavensis	2	7	0	3	1.000
Acp8	2	8	4	8	0.481
arizonae	1	4	----	----	----
mojavensis	1	4	----	----	----
Acp16a	0	11	2	7	0.189
arizonae	0	4	1	1	0.333
mojavensis	0	7	1	4	0.417
Acp16b	6	9	1	6	0.207
arizonae	3	7	1	4	0.675
mojavensis	3	2	0	0	----
Acp19	5	7	2	11	0.139
arizonae	3	4	2	7	0.377
mojavensis	3	3	0	4	0.200
Acp21a	1	11	2	21	0.971
arizonae	1	2	----	----	----
mojavensis	1	9	----	----	----
Acp24	2	6	1	1	0.504
arizonae	0	2	----	----	----
mojavensis	2	4	----	----	----
Acp25	8	2	2	6	0.017*
arizonae	6	1	1	2	0.103
mojavensis	3	1	1	3	0.148

Table 2.5. Contd.

Gene	Polymorphic		Fixed		P^a
	Syn	Repl	Syn	Repl	
Acp27a	2	5	0	1	1.000
arizonae	0	1	0	1	----
mojavensis	2	4	0	0	----
Acp42	7	6	3	11	0.078
arizonae	2	3	----	----	----
mojavensis	5	3	----	----	----
Acp48	7	9	14	30	0.396
arizonae	2	4	----	----	----
mojavensis	5	5	----	----	----
moj9	12	6	3	0	0.526
arizonae	6	4	1	0	1.000
mojavensis	6	2	1	0	1.000
moj30	21	10	3	0	0.539
arizonae	10	4	1	0	1.000
mojavensis	13	6	2	0	1.000
Tes14	3	0	0	0	----
arizonae	1	0	0	0	----
mojavensis	2	0	0	0	----
Tes33	24	7	3	0	0.589
arizonae	15	5	1	0	1.000
mojavensis	10	2	2	0	1.000
Tes100	3	7	1	1	0.592
arizonae	0	5	1	0	----
mojavensis	3	2	0	1	----
Tes101	1	1	1	0	----
arizonae	1	0	0	0	----
mojavensis	0	1	1	0	----
Tes104	14	2	7	0	0.557
arizonae	9	2	3	0	1.000
mojavensis	6	0	4	0	----
Tes105	4	4	0	0	----
arizonae	2	2	0	0	----
mojavensis	2	2	0	0	----
Tes106	6	4	5	0	0.231
arizonae	2	2	3	0	0.429
mojavensis	4	2	2	0	1.000

Table 2.5. Contd.

Gene	Polymorphic		Fixed		
	Syn	Repl	Syn	Repl	P^a
Tes107	5	0	1	0	----
arizonae	3	0	0	0	----
mojavensis	2	0	1	0	----
Tes109	3	10	1	4	0.887
arizonae	3	6	----	----	----
mojavensis	0	4	----	----	----
Tes110	2	3	6	0	0.061
arizonae	2	1	0	0	----
mojavensis	0	2	6	0	----
Tes112	7	0	0	1	----
arizonae	2	0	----	----	----
mojavensis	5	0	----	----	----
Tes113	3	3	2	1	0.633
arizonae	1	2	----	----	----
mojavensis	3	1	----	----	----
Tes114	1	0	1	0	----
arizonae	0	0	1	0	----
mojavensis	1	0	0	0	----
Tes115	0	3	2	2	0.429
arizonae	0	2	1	1	----
mojavensis	0	1	1	1	----
Tes118	6	8	2	4	0.688
arizonae	2	6	----	----	----
mojavensis	4	2	----	----	----
Tes134	9	5	5	2	0.742
arizonae	8	1	2	0	1.000
mojavensis	1	4	3	2	0.189
Tes154	9	3	4	0	0.529
arizonae	1	1	2	0	----
mojavensis	8	2	1	0	1.000

[a] *P*-values from G-tests, Fisher's exact test when zero values are present. Significant results are indicated by an asterisk. Tests were not carried out for loci with very few observations.

Table 2.6. McDonald-Kreitman Tests for Gene Classes

	Synonymous	Replacement	
	moj- genes		
Polymorphic	33	16	Fisher's exact test:
Fixed	6	0	$P = 0.165$
	All testis-expressed genes		
Polymorphic	100	60	$G = 2.162$
Fixed	41	15	$P = 0.142$
	All *Acps*		
Polymorphic	63	115	$G = 6.474$
Fixed	42	139	$P = 0.011$*
	All *Acps* except *Acp*25		
Polymorphic	55	113	$G = 3.91$
Fixed	40	133	$P = 0.047$*

Note.—Probability determined by a G-test when all cells contain non-zero values, Fisher's exact test otherwise.

Table 2.7. Polarized McDonald-Kreitman Tests for Gene Classes

	Synonymous	Replacement	
	D. mojavensis moj- genes		
Polymorphic	19	8	Fisher's exact test:
Fixed	3	0	$P = 0.545$
	D. mojavensis testis-expressed genes		
Polymorphic	39	18	$G = 2.295$
Fixed	21	4	$P = 0.130$
	D. mojavensis Acps		
Polymorphic	21	38	$G = 8.329$
Fixed	2	24	$P = 0.004$*
	D. arizonae moj- genes		
Polymorphic	16	8	Fisher's exact test:
Fixed	2	0	$P = 0.557$
	D. arizonae testis-expressed genes		
Polymorphic	44	21	$G = 4.967$
Fixed	14	1	$P = 0.026$*
	D. arizonae Acps		
Polymorphic	22	32	$G = 1.792$
Fixed	11	29	$P = 0.181$

Note.—Probability determined by a G-test when all cells contain non-zero values, Fisher's exact test otherwise.

Table 3.1. Sample and Distribution of Duplicate Genes

Duplicate Gene	Sample		Documented in the Same Fly Line?			
	ari	*moj*	*a* + *b*	*a* + *c*	*b* + *c*	*a* + *b* + *c*
Acp5a	7	7				
Acp5b	3	1	3 *ari,* 1 *moj*	1 *moj*	no	no
Acp5c	0	1				
Acp16a	7	6				
Acp16b	7	4	7 *ari,* 3 *moj*	2 *moj*	1 *moj*	no
Acp16c	0	3				
Acp21a	6	7	no	----	----	----
Acp21b	1	0				
Acp27a	7	7	5 *moj*	----	----	----
Aco27b	0	5				

Table 3.2. Polymorphism and Interspecific (Orthologous) Divergence of Duplicate *Acps*

Gene	No. alleles	Sample	Number Sites		$\theta_{syn.}$	$\theta_{rep.}$	Ks	Ka	Ka/Ks
			Syn	Repl					
Acp5a	7	*ari*	27	72	0.0151	0.0057	0.111	0.110	0.990
Acp5a	7	*moj*	27	72	0.0000	0.0170	----	----	----
Acp5b	3	*ari*	27	69	0.0000	0.0000	0.000	0.112	Ka>Ks
Acp16a	7	*ari*	38	103	0.0000	0.0159	0.060	0.132	2.205
Acp16a	6	*moj*	38	103	0.0000	0.0299	----	----	----
Acp16b	7	*ari*	49	155	0.0251	0.0184	0.062	0.050	0.808
Acp16b	4	*moj*	49	155	0.0336	0.0070	----	----	----
Acp16c	3	*moj*	45	156	0.0000	0.0086	----	----	----
Acp21a	6	*ari*	48	132	0.0092	0.0066	0.055	0.227	4.121
Acp21a	7	*moj*	48	132	0.0086	0.0278	----	----	----
Acp27a	7	*ari*	68	214	0.0000	0.0019	0.006	0.013	2.138
Acp27a	7	*moj*	68	214	0.0120	0.0076	----	----	----
Acp27b	5	*moj*	71	208	0.0068	0.0115	----	----	----
all Dupls.		*ari*	257	745	0.0080	0.0083	0.044	0.094	2.123
		moj	346	1040	0.0097	0.0139	----	----	----
other *Acps*		*ari*	712	2336	0.0149	0.0058	0.068	0.052	0.761
		moj	712	2336	0.0166	0.0075	----	----	----

Table 3.3. Intraspecific (Paralogous) Divergence of Duplicate *Acps*

Gene Pair	# alleles first Dupl.	# alleles second Dupl.	Number Sites		Ks	Ka	Ka/Ks
			Syn	Repl			
Acp5							
ari (a : b)	7	3	27	69	0.199	0.272	1.370
moj (a : b)	7	1	24	63	0.043	0.205	4.799
moj (a : c)	7	1	25	65	0.124	0.474	3.817
moj (b : c)	1	1	25	68	0.157	0.434	2.757
Acp16							
ari (a : b)	7	7	40	116	0.229	0.442	1.934
moj (a : b)	6	4	40	113	0.247	0.461	1.867
moj (a : c)	6	3	40	116	0.196	0.314	1.599
moj (b : c)	4	3	46	149	0.313	0.378	1.209
Acp21							
ari (a : b)	6	1	49	137	0.014	0.134	9.734
Acp27							
moj (a : b)	7	5	65	196	0.021	0.103	4.899

Table 3.4. Branch-Specific Divergence of *Acp21* and *Acp27* Duplicate Families

Gene Family	Ka	Ks	$2\Delta l$[a]
Acp21			
ari a	0.089	0.001	*6.15
ari b	0.059	0.001	*4.55
moj a	0.186	0.005	**6.78
Acp27			
ari a	0.014	0.000	1.790
moj a	0.000	0.013	----
moj b	0.110	0.000	**12.26

[a]likelihood-ratio tests vs. the null model (Ka = Ks), * indicates $P < 0.05$, ** indicates $P < 0.01$.

Table 3.5. McDonald-Kreitman Tests of Duplicate Gene Pairs

Gene (pair)	Polymorphic		Fixed		*P*	Outgroup
	Syn	Repl	Syn	Repl		
Acp5						
ari a / moj a	1	4	3	6	0.590	----
ari a	1	1	3	3	1.000	*ari b*
moj a	0	3	0	2	----	*ari b*
ari a / ari b	1	1	5	15	0.473	----
ari a	1	1	2	4	0.676	*moj c*
ari b	0	0	1	6	----	*moj c*
ari a / moj c	1	1	4	25	0.245	----
Acp16						
ari a / moj a	0	11	2	7	0.189	----
ari a	0	4	1	1	0.333	*moj c*
moj a	0	7	1	4	0.417	*moj c*
ari a / moj c	0	6	8	19	0.296	----
ari a	0	4	1	9	1.000	*moj b*
moj c	0	2	6	7	0.486	*moj b*
ari b / moj b	6	9	1	6	0.207	----
ari b	3	7	1	4	0.675	*moj c*
moj b	3	2	0	0	----	*moj c*
moj b / moj c	3	4	13	41	0.309	----
Acp21						
ari a / moj a	1	11	2	21	0.971	----
ari a / ari b	1	5	1	13	0.531	----
ari a	1	5	1	5	1.000	*moj a*
ari b	----	----	0	6	----	*moj a*
Acp27						
ari a / moj a	2	5	0	1	1.000	----
ari a	0	1	0	1	----	*moj b*
moj a	2	4	0	0	----	*moj b*
moj a / moj b	3	8	0	17	0.050	----
moj a	2	3	0	0	----	*ari a*
moj b	1	5	0	17	0.261	*ari a*

Table 4.1. Gene Intron/Exon Structure, Signal Peptide Prediction, and Amino Acid Sequence Identity Between *D. melanogaster* and *D. pseudoobscura Acps*

Gene	#AA residues	exon(s)[a]	intron(s)[a]	Signal peptide[b]	% Similar[c]
Acp26Aa	264	34, 761	56	1.00	18.5
pse-Acp26Aa	250	37, 716	68	1.00	
Acp26Ab	90	31, 242	61	1.00	33.3
pse-Acp26Ab	92	31, 248	65	1.00	
Acp29AB	234	705	----	0.99	
Acp32CD	252	759	----	0.99	43.7
pse-Acp32CD	299	900	----	1.00	
Acp33A	47	144	----	0.97	
Acp36DE	912	208, 2531	59	0.98	
Acp53Ea	120	42, 321	65	1.00	41.7
pse-Acp53Ea	120	42, 321	72	1.00	
Acp53C14a	121	42, 324	52	1.00	55.0
pse-Acp53C14a	120	42, 321	71	1.00	
Acp53C14b	132	42, 357	56	1.00	48.5
pse-Acp53C14b	132	42, 357	65	1.00	
Acp53C14c	124	42, 333	57	0.99	40.5
pse-Acp53C14c	121	42, 324	56	0.99	
pse-Acp53C14d	129	42, 348	70	1.00	
pse-Acp53C14e	127	42, 342	51	0.99	
pse-Acp53C14f	127	33, 351	60	1.00	

Table 4.1. Contd.

Gene	#AA residues	exon(s)[a]	intron(s)[a]	Signal peptide[b]	% Similar[c]
Acp62F	115	348	----	1.00	42.0
pse-Acp62F	135	408	----	0.99	
Acp63F	81	28, 156, 62	61, 54	1.00	
Acp70A	55	115, 53	65	1.00	54.7
pse-Acp70A	57	118, 56	74	1.00	
Acp95EF	52	18, 141	62	1.00	
Acp98AB	28-31	87-96	----	0.00	

[a]Number of nucleotides per exon/intron, starting from the initiation codon and going through the stop codon.

[b]Probability of signal peptide as predicted by the hidden Markov method of SignalP, version 3.0 (Nielsen and Krogh 1998; Bendtsen et al. 2004).

[c]Percent amino acid identities, calculated as the number of identical residues/total number of alignable residues.

Table 4.2. Accession Nos. and Initiation Codon Positions for *D. pseudoobscura Acp* Orthologs and Microsyntenic Contigs

Gene	Accession Nos.	position[a]	strand[b]
pse-Acp26Aa	AADE01000400	9279	-
pse-Acp26Ab	AADE01000400	7192	-
pse-Acp32CD	AADE01000037	191188	-
pse-Acp53Ea	AADE01000143	121103	-
pse-Acp53C14a	AADE01000143	119222	-
pse-Acp53C14b	AADE01000143	120132	-
pse-Acp53C14c	AADE01001461	25072	+
pse-Acp53C14d	AADE01000143	121785	-
pse-Acp53C14e	AADE01000143	122365	-
pse-Acp53C14f	AADE01000143	122911	-
pse-Acp62F	AADE01003187	3724	+
pse-Acp70A	AADE01000940	4090	+

Microsyntenic Region[c]	Accession Nos.
Acp29AB	AADE01000153
Acp33a proximal	AADE01004963
Acp33a distal	AADE01000551
Acp36DE	AADE01001378
Acp62F	AADE01001729
Acp63F	AADE01002121
Acp70A	AADE01003892
Acp76A	AADE01001646
Acp95EF	AADE01000038
Acp98AB	AADE01000028

[a] Nucleotide position of the first base of the start codon for *D. pseudoobscura Acps*.
[b] Indicates whether the *Acp* is on the plus or minus strand of the indicated contig.
[c]Accession Nos. are for *D. pseudoobscura* homologous regions corresponding to *D. melanogaster Acps* (see Figs. 4.6-4.12). There are two *D. pseudoobscura* accessions for the *Acp33A* region, due to incomplete genome assembly (see Fig. 4.7).

Table 4.3. Silent and Replacement Polymorphism and Divergence for *Acp26Aa* in *D. melanogaster* and *D. pseudoobscura*

Sample	No. of Sites		Sil Theta	Rep Theta	Ks[a]	Ka[a]	Ka/Ks
	Sil.	Repl.					
pse[b]	154	524	0.034	0.008	0.096	0.100	1.038
pse + *per*[c]	154	524	0.037	0.010	0.097	0.101	1.034
mel (USA)[d]	174	615	0.014	0.006	0.167	0.156	0.934
mel (Malawi)[d]	174	615	0.033	0.008			

[a]Divergence estimates are with respect to *D. miranda* and *D. simulans* for *D. pseudoobscura*/*D. persimilis* and *D. melanogaster*, respectively.

[b]Population genetic data are restricted to the six *D. pseudoobscura* alleles.

[c]Population genetic data includes the six *D. pseudoobscura* alleles as well as a single *D. persimilis* allele.

[d]*D. melanogaster* polymorphism data are from Aguadé (1998). *D. melanogaster* divergenge data are from Aguadé et al. (1992).

Table 4.4. McDonald-Kreitman tests of neutral molecular evolution at *Acp26Aa* in *D. melanogaster* and *D. pseudoobscura*

Sample	Polymorphic		Fixed[a]		P[b]
	Syn	Repl	Syn	Repl	
pse[c]	10	9	12	39	0.022
pse + *per*[c]	12	11	12	39	0.016
mel (USA)[d]	7	9	24	78	0.109
mel (Malawi)[d]	19	15	20	77	0.002

[a]Fixations are with respect to *D. miranda* and *D. simulans* for *D. pseudoobscura*/*D. persimilis* and *D. melanogaster*, respectively.

[b]Probability determined by G-test.

[c]Polymorphism and fixation data as calculated by excluding (*pse*) and including (*pse* + *per*) the single *D. persimilis* allele.

[d]*D. melanogaster* data are from Aguadé (1998).

References

Aguadé, M. 1997. Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*. Mol. Biol. Evol. **14**:544-549.

Aguadé, M. 1998. Different forces drive the evolution of the *Acp26Aa* and *Acp26Ab* accessory gland genes in the *Drosophila melanogaster* species complex. Genetics **150**:1079-1089.

Aguadé, M. 1999. Positive selection drives the evolution of the *Acp29AB* accessory gland protein in *Drosophila*. Genetics **152**:543-551.

Aguadé, M., N. Miyashita, and C. H. Langley. 1992. Polymorphism and divergence in the *Mst26A* male accessory gland gene region in *Drosophila*. Genetics **132**:755-770.

Aigaki, T., I. Fleischmann, P. S. Chen, and E. Kubli. 1991. Ectopic expression of sex peptide alters reproductive behavior of female *D. melanogaster*. Neuron **4**:557-563.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389-3402.

Andrews, J., G. G. Bouffard, C. Cheadle, J. Lü, K. G. Becker and B. Oliver. 2000. Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. Genome Res. **10**:2030-2043.

Aquadro, C. F., K. M. Lado, and W. A. Noon. 1988. The *rosy* region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. Genetics **119**:875-888.

Arnqvist, G., M. Edvardsson, U. Friberg, and T. Nilsson. 2000. Sexual conflict promotes speciation in insects. Proc. Natl. Acad. Sci. **97**:10460-10464.

Ashburner, M. 1989. *Drosophila*: A laboratory handbook. Cold Spring Harbor Press, Cold Spring Harbor, N.Y.

Barbash, D.A., D.F. Siino, A.M. Tarone, and J. Roote. 2003. A rapidly evolving MYB-related protein causes species isolation in *Drosophila*. Proc. Natl. Acad. Sci. **100**:5302-5307.

Batzoglou, S., L. Pachter, J.P. Mesirov, B. Berger, and E.S. Lander. 2000. Human and mouse gene structure: comparative analysis and application to exon prediction. Genome Res. **10**:950-958.

Beckenbach A. T., Y. W. Wei and H. Liu. 1993. Relationships in the *Drosophila obscura* species group, inferred from mitochondrial cytochrome oxidase II sequences. Mol. Biol. Evol. **10**:619-634.

Begun, D. J. 1996. Population genetics of silent and replacement variation in *Drosophila simulans* and *D. melanogaster*: X/autosome differences? Mol. Biol. Evol. **13**:1405-1407.

Begun, D. J., and P. Whitley. 2002. Molecular population genetics of *Xdh* and the evolution of base composition in *Drosophila*. Genetics **162**:1725-1735.

Begun, D. J., P. Whitley, B. L. Todd, H. M. Waldrip-Dail and A. G. Clark. 2000. Molecular population genetics of male accessory gland proteins in *Drosophila*. Genetics **156**:1879-1888.

Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. **340**:783-795.

Bergman, C. M., B. D. Pfeiffer, D. E. Rincón-Limas et al. (17 co-authors). 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. Genome Biol. **3**: research0086.1-0086.20.

Bernasconi, G., T. L. Ashman, T. R. Birkhead et al. (15 co-authors). 2004. Evolutionary ecology of the prezygotic stage. Science **303**:971-975.

Bertram, M. J., D. M. Neubaum and M. F. Wolfner. 1996. Localization of the *Drosophila* accessory gland protein *Acp36DE* in the mated female suggests a role in sperm storage. Insect Biochem. Molec. Biol. **26**:971-980.

Betrán, E., and M. Long. 2003. *Dntf-2r*, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. Genetics **164**:977-988.

Betrán, E., K. Thornton, and M. Long. 2002. Retroposed new genes out of the X in *Drosophila*. Genome Res. **12**:1854-1859.

Birkhead, T. R., and A. P. Møller. 1998. Sperm competition and sexual selection. Academic, London.

Birkhead, T. R., and T. Pizzari. 2002. Postcopulatory sexual selection. Nature Rev. Genet. **3**:262-273.

Caracristi, G., and C. Schlötterer. 2003. Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. Mol. Biol. Evol. **20**:792-799.

Carlson, J.R., and D.S. Hogness. 1985. The Jonah genes: a new multigene family in *Drosophila melanogaster*. Dev. Biol. **108**:341-354.

Chapman, T. 2001. Seminal fluid-mediated fitness traits in *Drosophila*. Heredity **87**:511-521.

Chapman, T., J. Bangham, G. Vinti, B. Seifried, O. Lung, M. F. Wolfner, H. K. Smith, and L. Partridge. 2003. The sex peptide of *Drosophila melanogaster*: Female post-mating responses analyzed by using RNA interference. Proc. Natl. Acad. Sci. **100**:9923-9928.

Chapman, T., J. Hutchings, and L. Partridge. 1993. No reduction in the cost of mating for *Drosophila melanogaster* females mating with spermless males. Proc. R. Soc. Lond. B. Biol. Sci. **253**:211-217.

Chapman, T., D. M. Neubaum, M. F. Wolfner, and L. Partridge. 2000. The role of male accessory gland protein *Acp36DE* in sperm competition in *Drosophila melanogaster*. Proc. R. Soc. Lond. B. Biol. Sci. **267**:1097-1105.

Chapman, T., L. F. Liddle, J. M. Kalb, M. F. Wolfner, and L. Partridge. 1995. Cost of mating in *Drosophila melanogaster* females is mediated by male accessory gland products. Nature **373**:241-244.

Chen, P. S., E. Stumm-Zollinger, T. Aigaki, J. Balmer, M. Bienz and P. Bohlen. 1988. A male accessory gland peptide that regulates reproductive behavior of female *D. melanogaster*. Cell **54**:291-298.

Clark, A. G., and D. J. Begun. 1998. Female genotypes affect sperm displacement in *Drosophila*. Genetics **149**: 1487-1493.

Clark, A. G., M. Aguadé, T. Prout, L. G. Harshman, and C. H, Langley. 1995. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics **139**:189-201.

Coulthart, M. B., and R. S. Singh. 1988. High level of divergence of male-reproductive-tract proteins, between *Drosophila melanogaster* and its sibling species, *D. simulans*. Mol. Biol. Evol. **5**:182-191.

Coulthart, M. B., and R. S. Singh. 1988. Differing amounts of genetic polymorphism in testes and male accessory glands of *Drosophila melanogaster* and *D. simulans*. Biochem. Genet. **26**: 153-164.

DiBenedetto, A. J., D. M. Lakich, W. D. Kruger, J. M. Belote, B. S. Baker and M. F. Wolfner. 1987. Sequences expressed sex-specifically in *Drosophila melanogaster* adults. Dev. Biol. **119**:242-251.

Dimarcq, J. -L., D. Hoffmann, M. Meister, P. Bulet, R. Lanot, J. -M. Reichhart, and J. A. Hoffmann. 1994. Characterization and transcriptional profiles of a *Drosophila* gene encoding an insect defensin: a study in insect immunity. Eur. J. Biochem. **221**:201-209.

Drysdale, R. 2003. The *Drosophila melanogaster* genome sequencing and annotation projects: a status report. Brief Funct. Genomic Proteomic **2**:128-134.

Eberhard, W. G. 1996. Female control: sexual selection by cryptic female choice. Princeton University Press, Princeton, NJ.

Eddy, E. M. 1998. Regulation of gene expression during spermatogenesis. Semin. Cell Dev. Biol. **9**:451-457.

Fry, C. L., and G. S. Wilkinson. 2004. Sperm survival in female stalk-eyed flies depends on seminal fluid and meiotic drive. Evolution Int. J. Org. Evolution. **58**:1622-1626.

Fuller, M. T. 1993. Spermatogenesis. Pp. 1-70 *in* M. Bate and A. Martinez-Arias, eds. The development of *Drosophila melanogaster*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Galvani, A.P., and M. Slatkin. 2003. Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele. Proc. Natl. Acad. Sci. **100**:15276-15279.

Gavrilets, S. 2000. Rapid evolution of reproductive barriers driven by sexual conflict. Nature **403**:886-889.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725-736.

Harshman, L. G., and T. Prout. 1994. Sperm displacement without sperm transfer in *Drosophila melanogaster*. Evolution **48**:758-766.

Heifetz, Y., and M. F. Wolfner. 2004. Mating, seminal fluid components, and sperm cause changes in vesicle release in the *Drosophila* female reproductive tract. Proc. Natl. Acad. Sci. **101**:6261-6266.

Heifetz, Y., O. Lung, E. A. Frongillo Jr., and M. F. Wolfner. 2000. The *Drosophila* seminal fluid protein *Acp26Aa* stimulates release of oocytes by the ovary. Curr. Biol. **10**:99-102.

Herndon, L. A., and M. F. Wolfner. 1995. A *Drosophila* seminal fluid protein, *Acp26Aa*, stimulates egg laying in females for 1 day after mating. Proc. Natl. Acad. Sci. **92**:10114-10118.

Hey, J., and R. Nielsen. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics **167**:747-760.

Holloway, A. and D. J. Begun. 2004. Molecular evolution and population genetics of duplicated accessory gland protein genes in *Drosophila*. Mol. Biol. Evol. **21**:1625-1628.

Jaillon, O., C. Dossat, R. Eckenberg et al. (11 co-authors). 2003. Assessing the *Drosophila melanogaster* and *Anopheles gambiae* genome annotations using genome-wide sequence comparisons. Genome Res. **13**:1595-1599.

Kalb, J. M., A. J. DiBenedetto, and M. F. Wolfner. 1993. Probing the function of *Drosophila melanogaster* accessory glands by directed cell ablation. Proc. Nalt. Acad. Sci. **90**:8093-8097.

Kern, A.D., C.D. Jones, and D.J. Begun. 2004. Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. Genetics **167**:725-735.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, UK.

Knowles, L. L., and T. A. Markow. 2001. Sexually antagonistic coevolution of a postmating-prezygotic reproductive character in desert *Drosophila*. Proc. Natl. Acad. Sci. **98**:8692-8696.

Krylov, D.M., Y.I. Wolf, I.B. Rogozin, and E.V. Koonin. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. **13**:2229:2235.

Kortschak, R.D., G. Samuel, R. Saint, and D.J. Miller. 2003. EST analysis of the cnidarian *Acropora millepora* reveals extensive gene loss and rapid sequence divergence in the model invertebrates. Curr. Biol. **13**:2190-2195.

Lakovaara, S., and A. Saura. 1982. Evolution and speciation in the *Drosophila obscura* group. Pp. 2-59 *in* M. Ashburner, H.L. Carson, and J.N. Thompson, Jr., eds. The genetics and biology of *Drosophila*, Vol. 3b. Academic Press, New York.

Lawniczak, M. K. N., and D. J. Begun. 2004. A genome-wide analysis of courting and mating responses in *Drosophila melanogaster* females. Genome **47**:1-11.

Li, W. 1995. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Liu, H., and E. Kubli. 2003. Sex-peptide is the molecular basis of the sperm effect in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. **100**:9929-9933.

Livak, K. J., and T. D. Schmittgen. 2001. Analysis of relative gene expression using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. Methods **25**:402-408.

Long, M., and C.H. Langley. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science **260**:91-95.

Long, M., E. Betrán, K. Thornton, and W. Wang. 2003. The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. **4**:865-875.

Lung, O., U. Tram, C. M. Finnerty, M. A. Eipper-Mains, J. M. Kalb, and M. F. Wolfner. 2002. The *Drosophila melanogaster* seminal fluid protein *Acp62F* is a protease inhibitor that is toxic upon ectopic expression. Genetics **160**:211-224.

Lynch, M., and A. Force. 2000. The origin of interspecific genomic incompatibility via gene duplication. Am. Nat. **156**:590-605.

Lynch, M., M. O'Hely, B. Walsh, and A. Force. 2001. The probability of preservation of a newly arisen gene duplicate. Genetics **159**:1789-1804.

Marchler-Bauer, A., J.B. Anderson, C. DeWeese-Scott et al. (27 co-authors). 2003. CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res. **31**:383-387.

Markow, T. A. 1982. Mating systems of cactophilic *Drosophila*. Pp. 273-287 *in* J. S. F. Barker and W. T. Starmer, eds. Ecological genetics and evolution: the cactus-yeast-*Drosophila* model system. Plenum Press, New York.

Markow, T. A. 1996. Evolution of *Drosophila* mating systems. Evol. Biol. **29**:73-106.

Markow, T. A. 2002. Perspective: female remating, operational sex ratio, and the arena of sexual selection in *Drosophila* species. Evolution **56**:1725-1734.

Markow, T. A., and P. F. Ankney. 1984. *Drosophila* males contribute to oogenesis in a multiple mating species. Science **224**:302-303.

Markow, T. A., and P. F. Ankney. 1988. Insemination reaction in *Drosophila*: found in species whose males contribute material to oocytes before fertilization. Evolution **42**:1097-1101.

Matzkin, L., and W. F. Eanes. 2003. Sequence variation of alcohol dehydrogenase (*Adh*) paralogs in cactophilic *Drosophila*. Genetics **163**:181-194.

McDonald, J. M., and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351**:652-654.

McGraw, L. A., G. Gibson, A. G. Clark, and M. F. Wolfner. Genes regulated by mating, sperm, or seminal proteins in mated female *Drosophila melanogaster*. 2004. Curr. Biol. **14**:1509-1514.

Meiklejohn, C. D., J. Parsch, J. M. Ranz, and D. L. Hartl. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. Proc. Natl. Acad. Sci. **100**:9894-9899.

Metz, E. C., and S. R. Palumbi. 1996. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. Mol. Biol. Evol. **13**:397-406.

Meyer, I. M., and R. Durbin. 2004. Gene structure conservation aids similarity based gene prediction. Nucleic Acids Res. **32**:776-783.

Moran, N.A. 2003. Tracing the evolution of gene loss in obligate bacterial symbionts. Curr. Opin. Microbial. **6**:512-518.

Moriyama, E. N., and J. R. Powell. 1996. Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13**:261-277.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.

Neubaum, D. M., and M. F. Wolfner. 1999. Mated *Drosophila melanogaster* females require a seminal fluid protein, *Acp36DE*, to store sperm efficiently. Genetics **153**:845-857.

Nielsen, H., and A. Krogh. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. Pp. 122-130 *in* Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 6). AAAI Press, Menlo Park, California.

Nurminsky, D. I., M. V. Nurminskaya, D. De Aguiar, and D. L. Hartl. 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature **396**:572-575.

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, Berlin.

Ohta, T. 1994. Further examples of evolution by gene duplication revealed through DNA sequence comparisons. Genetics **138**:1331-1337.

Olson, M. V. 1999. When less is more: gene loss as an engine of evolutionary change. Am. J. Hum. Genet. **64**:18-23.

Olson, M. V., and A. Varki. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. Nat. Rev. Genet. **4**:20-28.

Parisi, M., R. Nuttall, D. Naiman, G. Bouffard, J. Malley, J. Andrews, S. Eastman, and B. Oliver. 2003. Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science **299**:697-700.

Parker, G. A., and L. Partridge. 1998. Sexual conflict and speciation. Philos. Trans. R. Soc. Lond. B Biol. Sci. **353**:261-274.

Patterson, J. T. 1946. A new type of isolating mechanism in *Drosophila*. Proc. Natl. Acad. Sci. **32**:202-208.

Patterson, J. T. 1947. The insemination reaction and its bearing on the problem of speciation in the *mulleri* subgroup. Univ. Texas Publ. **4720**:41-77.

Patterson, J. T., and W. S. Stone. 1952. Evolution in the genus *Drosophila*. Macmillan, New York.

Pitnick, S., T. A. Markow, and G. S. Spicer. 1995. Delayed male maturity is a cost of producing large sperm in *Drosophila*. Proc. Natl. Acad. Sci. **92**:10614-10618.

Pitnick, S., T. A. Markow, and G. S. Spicer. 1999. Evolution of multiple kinds of female sperm-storage organs in *Drosophila*. Evolution **53**:1804-1822.

Pitnick, S., G. T. Miller, K. Schneider, and T. A. Markow. 2003. Ejaculate-female coevolution in *Drosophila mojavensis*. Proc. R. Soc. Lond. B **270**:1507-1512.

Pitnick, S., G. S. Spicer, and T. A. Markow. 1997. Phylogenetic examination of female incorporation of ejaculate in *Drosophila*. Evolution **51**:833-845.

Poccia, D. 1994. Molecular aspects of spermatogenesis. R. G. Landes Compay, Austin, TX.

Powell, J. R. and R. DeSalle. 1995. *Drosophila* molecular phylogenies and their uses. Evol. Biol. **28**:87-138.

Presgraves, D.C., L. Balagopalan, S.M. Abmayr, and H.A. Orr. 2003. Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. Nature **423**:699-700.

Pyle, D. W., and M. H. Gromko. 1981. Genetic basis for repeated mating in *Drosophila melanogaster*. Am. Nat. **117**:133-146.

Ranz, J. M., C. I. Castillo-Davis, C. D. Meiklejohn, and D. L. Hartl. 2003. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science **300**:1742-1745.

Reed, L. K., and T. A. Markow. 2004. Early events in speciation: polymorphism for hybrid male sterility in *Drosophila*. Proc. Natl. Acad. Sci. **101**:9009-9012.

Rice, W. R. 1996. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature **381**:232-234.

Rice, W. R. 1998. Intergenomic conflict, interlocus antagonistic coevolution, and the evolution of reproductive isolation. Pp. 261-270 *in* D. J. Howard and S. H. Berlocher, eds. Endless forms: species and speciation. Oxford University Press, New York.

Rifkin, S. A., J. Kim, and K. P. White. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nat. Genet. **33**:138-144.

Rozas, J., and R. Rozas. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15**:174-175.

Saudan, P., K. Hauck, M. Soller et al. (12 co-authors). 2002. Ductus ejaculatorius peptide 99B (*DUP99B*), a novel *Drosophila melanogaster* sex-peptide pheromone. Eur. J. Biochem. **269**:989-997.

Schäfer, U. 1986. Genes for male-specific transcripts in *Drosophila melanogaster*. Mol. Gen. Genet. **202**:219-225.

Singh, S. R., B. N. Singh, and H. F. Hoenigsberg. 2002. Female remating, sperm competition and sexual selection in *Drosophila*. Genet. Mol. Res. **1**:178-215.

Sorhannus, U. 2003. The effect of positive selection on a sexual reproduction gene in *Thalassiosira weissflogii* (Bacillariophyta): results obtained from maximum-likelihood and parsimony-based methods. Mol. Biol. Evol. **20**:1326-1328.

Steinemann, M., W. Pinsker, and D. Sperlich. 1984. Chromosome homologies within the *Drosophila Obscura* group. Chromosoma **91**:46-53.

Stevison, L. S., B. A. Counterman, and M. A. F. Noor.2004. Molecular evolution of X-linked accessory gland proteins in *Drosophila pseudoobscura*. J. Hered. **95**:114-118.

Sutton, K. A., and M. F. Wilkinson. 1997. Rapid evolution of a homeodomain: evidence for positive selection. J. Mol. Evol. **45**:579-588.

Swanson, W. J. and V. D. Vacquier. 1995. Extraordinary divergence and positive Darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. Proc. Natl. Acad. Sci. **92**:4957-4961.

Swanson, W. J. and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat. Rev. Genet. **3**:137-144.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner and C. F. Aquadro. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc. Natl. Acad. Sci. USA **98**:7375-7379.

Ting, C.T., S.C. Tsaur, M.L. Wu, and C.I. Wu. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. Science **282**:1501-1504.

Torgerson, D. G., R. J. Kulathinal, and R. S. Singh. 2002. Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. Mol. Biol. Evol. **19**:1973-1980.

Tram, U., and M. F. Wolfner. 1999. Male seminal fluid proteins are essential for sperm storage in *Drosophila melanogaster*. Genetics **153**:837-844.

Tsaur, S. C., and C. I. Wu. 1997. Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*. Mol. Biol. Evol. **14**:544-549.

Tsaur, S. C., C. T. Ting, and C. I. Wu. 1998. Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa*, of *Drosophila* II. Divergence versus polymorphism. Mol. Biol. Evol. **15**:1040-1046.

Vacquier, V. D. 1998. Evolution of gamete recognition proteins. Science **281**:1995-1998.

Walsh, J. B. 1995. How often do duplicated genes evolve new functions? Genetics **139**:421-428.

Watanabe, T. K. and M. Kawanishi. 1979. Mating preference and the direction of evolution in Drosophila. Science. **205**:906-907.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**:256-276.

Wheeler, M. R. 1947. The insemination reaction in intraspecific matings of *Drosophila*. Univ. Texas Publ. **4720**:78-115.

Wiehe, T., S. Gebauer-Jung, T. Mitchell-Olds, and R. Guigó. 2001. SGP-1: Prediction and validation of homologous genes based on sequence alignments. Genome Res. **11**:1574-1583.

Wolfner, M. F. 1997. Tokens of love: functions and regulation of *Drosophila* male accessory gland products. Inst. Biochem. Mol. Biol. **27**:179-192.

Wolfner, M. F. 2002. The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. Heredity **88**:85-93.

Wolfner, M. F., H. A. Harada, M. J. Bertram, T. J. Stelnick, K. W. Kraus, J. M. Kalb, Y. O. Lung, D. M. Neubaum, M. Park and U. Tram. 1997. New genes for male accessory gland proteins in *Drosophila melanogaster*. Insect Biochem. Molec. Biol. **27**:825-834.

Wright, F. 1990. The "effective number of codons" used in a gene. Gene **87**:23-29.

Wyckoff, G. J., W. Wang, and C.I. Wu. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature **503**:304-309.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555-556 (http://abacus.gene.ucl.ac.uk/software/paml.html).

Yang, Z 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. **15**:568-573.

Zouros, E. 1973. Genic differentiation associated with the early stages of speciation in the *mulleri* subgroup of *Drosophila*. Evolution **27**:601-621.

Vita

Bradley Jon Wagstaff was born in Coral Gables, Florida on February 15, 1970, the son of Karma Lynn Wagstaff and Ronald Albert Wagstaff. After completing his work at Salmen High School, Slidell, Louisiana, in 1988, he entered Brigham Young University in Provo, Utah. He transferred to the University of New Orleans in August 1990, where he received the degree of Bachelor of Sciences in May 1996. In September 1996 he entered the Graduate School of the University of Texas at Austin in the Department of Zoology.

Permanent Address: 4137 Palmyra St., New Orleans, Louisiana 70119

This dissertation was typed by the author.